

‘An Eye for an Aye’: Linguistic and Political Backlash and Conformity in Eighteenth-Century Scots

LING690

Sarah van Eyndhoven

Abstract

This study examines the effects of social and political changes that were occurring during the eighteenth century in Scotland on the use of written Scots, focussing in particular upon authors who were known to have been for or against the Union of the Parliaments in 1707. In order to capture a holistic representation of the levels of Scots in writing, I explore the proportion of Scots lexemes, compared with their corresponding English lexemes, in a purpose-built corpus containing a range of eighteenth-century texts. This corpus contains both texts that were produced by a general cross-section of Scottish society, and a number of politically-active individuals. I take a quantitative sociolinguistic approach to historical data by utilising statistical techniques that examine linguistic variation in a data-driven manner. This enables a more detailed and empirical exploration of Scots in the eighteenth century, which until now has been largely examined on a descriptive basis only. Using a number of statistical tools that are well suited to historical analyses, such as Variability-based Neighbour Clustering (Gries & Hilpert, 2008), conditional inference trees (Hothorn et al., 2006) and random forests (Breiman, 2001), I have been able to reconstruct both the general patterning of the Scots language over time and the extralinguistic factors encouraging or suppressing its presence in writing. In particular, I compare the use of Scots between the general literate population and political individuals active during this time period. I also explore the effect of the latter’s political sympathies on their language choices, and uncover several new and interesting effects conditioning the levels of Scots in their writings. I tie these results to the underlying political change and discontent characterising Scotland during this time, as well as the general linguistic changes taking place across the eighteenth century as a result of broader processes of change over time.

Contents

1.0	Introduction	4
2.0	Literature Review	11
2.1	Historical Background - Linguistic Change	11
2.1.1	The Union of the Parliaments, 1707	11
2.1.2	The Age of Politeness	13
2.1.3	Scottish Standard English	17
2.1.4	Antiquarianism	19
2.1.5	Vernacular Revival	22
2.2	Historical Background - Political Change	26
2.2.1	Political Tension	26
2.2.2	Radicalism	30
2.3	Divergence and Convergence	35
2.3.1	Intertwining Influences	37
2.3.2	The retainment of Scots	39
2.3.3	The complexities of the eighteenth century	41
2.4	Previous Analyses	41
3.0	Research Questions	44
4.0	Methodology	45
4.1	The Corpus	45
4.1.1	Corpus Compilation	45
4.1.2	OCR	55
4.1.3	LaBB-CAT	58
4.2	The variable phenomenon	67
4.2.1	Dictionary compilation	68
4.2.2	Extracting the variable	86
4.2.3	Circumscribing the data	87
4.2.4	Recoding factor levels	88
4.3	The data process revisited	92
5.0	Results and Discussion	94
5.1	The benefits of Statistical Modelling	94
5.2	Statistical Modelling – The Temporal Analysis	96
5.2.1	Variability-Based Neighbour Clustering (VNC)	96
5.2.2	VNC - The general literate Scottish population	97
5.2.3	VNC – the political members of the corpus	102
5.3	Statistical Modelling – The Extralinguistic Factors	105

5.3.1 Conditional Inference Trees	105
5.3.2 Random Forests	108
5.3.3 General Scottish Society	112
5.3.4 Politically-Active Scottish Society	130
6.0 Further Discussion.....	147
7.0 Limitations and Future Directions.....	153
8.0 Conclusion	155
9.0 References	157
Acknowledgements.....	175
Appendices.....	177
Appendix One.....	177
Appendix Two	178
Appendix Three	178
List of Figures	180

1.0 Introduction

The relationship between language and politics has been attested in a number of studies (Hall-Lew et al., 2017, Hall-Lew et al., 2010, Kirkham and Moore, 2016), though many are focussed largely on contemporary or fairly recent cases of political change or politicians. These studies are often based around language manipulation for particular political purposes, or the language of politicians ascribing to certain political identities. Hall-Lew et al. (2017) analysed a number of marked phonological variants in speakers from the main Scottish political parties in current-day Scotland. They found a suggestive main effect for political party determining vowel height, indicating that variation indexes political identity. Politicians from the Scottish Labour Party, who believe Scotland should remain part of the United Kingdom (Hassan & Shaw, 2012) produced a higher ‘CAT’ vowel (Johnston, 1997), than members from the Scottish National Party (SNP), who wish Scotland to become independent. Hall-Lew et al. (2017) argue that a higher CAT vowel is linked to a middle class, conservative persona, and accordingly members of the SNP, being inherently anti-institutional, produce a lower CAT vowel overall. This indicates an interaction between political affiliation and linguistic choices, at least for members of the modern-day Scottish parliament. The language of historical Scottish politicians, on the other hand, has not yet been examined. The situation of the modern, devolved Scottish parliament, based in Scotland, is somewhat different to that of the unified British parliament of the eighteenth century, based at Westminster (see section 2.1.1), and so it is unclear whether the effect of political ideology found by Hall-Lew et al. (2017) might be comparable to eighteenth-century politicians. However, it is plausible that similar loyalties existed in historical times, which could in turn influence language usage.

There has been some recognition of the link between political change and language use across language communities in a number of historical cases. Usually this manifests as patriotic resistance to an unwelcome political regime that encroaches upon the local vernacular. This occurred for instance in Catalonia, which saw the continued use of Catalan during the Franco dictatorship,

despite its use being banned (Joseph, 2006). Similarly, Finnish went underground and continued to be spoken during the Swedish and later Russian conquests (McClure, 1980). The continuation of the Welsh language, following the 1536 Act of the Union between England and Wales, is a clear example of linguistic backlash by a local population in reaction to enforced political domination. In response to the repressive Act, which did not grant Welsh official language status and sought to exclude it from most higher-state functions, the Welsh language took on new importance as a keystone of Welsh identity (Phillips, 2012). Furthermore, in a bid to disseminate the King's religion throughout Wales, the Act established a Welsh translation of the Bible and Book of Common Prayer, which became standard household literature. Thus, rather than eroding the existing language structure, the legislation served only to promote the vernacular, further strengthening the function of Welsh as a marker of national identity (Phillips, 2012).

Although Scotland and England were similarly joined by a Union (or rather, two Unions to be exact)¹ that saw Scots increasingly lose its position in many high-level functions, the effects of this on the population at large, as well as those involved in transacting the agreement, are not particularly clear. There has been very little research on any interaction between politics and historical Scots in general, although Robinson (1983) did consequentially consider these factors in her descriptive account of scribal language use during the Reformation. Robinson (1983) analysed scribal versions of the Scots Confession that were created during the Reformation of 1560, which saw Scotland's established religion change from Catholicism to Protestantism. The texts indicated large amounts of variation and a lack of strong linguistic attitudes, although scribes seeking Protestant support frequently chose English options over Scottish, in order to produce a comprehensible English text. Overall however, there does not seem to have been any particular linguistic preference throughout,

¹ Scotland and England were joined twice by a Union. The first Union occurred in 1604, known as the Union of the Crowns. This saw the two nations come under one monarchy, under the Scottish King James VI. However, the Scottish Parliament, as well as various legal, educational and ecclesiastical structures remained separate. The second Union occurred in 1707; this was the Union of the Parliaments, which formally dissolved Scotland's independent parliament and instead incorporated a number of Scottish politicians into the one, overarching parliament at Westminster.

and Robinson (1983) suggests there was little interaction between politico-religious events and language use in Old Scots.

However, the focus was on a small subset of texts, with no ability to utilise today's statistical tools to examine the various, potentially interacting factors. Indeed, other accounts have suggested links between the religious controversies of the Reformation and the linguistic choices of individuals involved (Dossena, 2000, 2009). For instance, John Knox (1513 – 1572) was criticised for 'knapping Suddroun' (to speak Southern English in a mincing or affected way; www.dsl.ac.uk), indicating an awareness of anglicised speech, and its ties to political-religious identities already well before the treaties uniting Scotland and England took place. Furthermore, Robinson's research examined the language of a few select individuals, producing a particular genre of text that was itself highly stylised and codified in its design. This is thus unable to demonstrate how the language of those driving the change, as well as that of the general population, is affected by a major, nationwide political/religious change taking place.

Although accounts specifically examining political change and language change in historical Scots are rare, those focussed upon eighteenth-century Scots have identified changes taking place both in linguistic attitudes (Jones, 1995; Millar, 2004) and in creative literature following the wake of the Union of 1707 between Scotland and England (Robinson, 1973; Dossena, 1997; Smith, 1996, 2007; Corbett, 2013). Such changes are at least indicative of the political influence. The linguistic after-effects of the Union upon elite society, and especially the 'vernacular revival' and 'patriotic backlash' of Scots that flourished in late eighteenth-century poetry, has been mentioned in numerous accounts (Clive, 1970; Robinson, 1973; McClure, 1980; Beal, 1997; Jones, 1997; Corbett et al., 2003; Corbett, 2013). In particular, there has been significant attention given to the language of the celebrated poets Allan Ramsay (1686 – 1758), Robert Fergusson (1750 – 1774) and Robert Burns (1759 – 1796), who chose to write in the language of their local vernacular (Robinson, 1973; Dossena, 2013; Smith, 1996, 2007; Mathison, 1995, 2007; Corbett, 2013). These studies provide

insight into the language use and linguistic flourishes of two exceptional individuals in a highly creative sphere, who were known to have been actively resisting the anglicisation of Scots, but they do not tell us a lot about what the rest of the literate Scottish society was doing. The focus is again very narrow and largely on the linguistic choices present in a small sub-selection of texts. We cannot compare these patterns with the rest of the literate Scots population, but the high level of manual analysis and very nature of such concise studies prohibits extending the scope to a larger range of texts.

Early Modern Scots has seen very little in the way of quantitative research at all, and corpus-based analysis has remained infrequent also. The few studies quantifying eighteenth-century Scots usage tend to be restricted in their scope, focussing on one or a few particular orthographic variables or lexical items across a small collection of texts (Cruickshank, 2012, 2017). Studies such as Corbett's (2013) analysis of the poems of Ramsay and Burns have tried to examine a wider range of Scots features, but compensate for this by analysing a limited number of texts (in Corbett's case the analysis was restricted to just two poems). It is unclear whether the factors argued to influence the variation can be applied to a greater pool of texts.

The tendency to rely on manual analysis negates the possibility to analyse a wider range of texts, whilst corpus-based studies of Old and Middle Scots, (see Meurman-Solin, 1989a, 1989b, 1989c, 1992, 1993a, 1997a, 2000b, 2003a) have focussed on a select few features – for example Devitt (1989a) looked at five Scottish orthographic and syntactic variants (the morphemes <-ing>, <-ed> and <wh-> and the lexical variants *no/nae* and *a/ane*). Such analyses of individual features can provide interesting insights into the trajectory and influences shaping the path of a particular variant, but we cannot acquire an overall picture of what the Scots language was doing at a given moment, and our knowledge of Scots and its changes is somewhat piecemeal as a result. The result is a lack of quantitative studies examining the potential language-politics interaction in eighteenth century Scots writing, especially across a broader range of material, and a larger section of society.

However, advances in the last ten years have made large-scale, empirically-robust analyses of historical data much more attainable.

In particular, there has been increasing recognition within the fields of Historical English linguistics and historical sociolinguistics, of the benefits of using contemporary statistical methods to quantitatively examine diachronic data (see for example Gries & Hilpert, 2008, 2010, 2012; Gries, 2016). Within Historical Scots the trend has been slower to catch on (although see Smith, (forthcoming)), and despite some early innovations (e.g. Romaine, 1982), current research methodologies have not kept up with recent advances. Most previous quantitative work has tended to focus on factors in isolation, rather than examining the effect of multiple predictors (see Devitt, 1989a; Meurman-Solin, 1989a, 1989b, 1989c, 1992, 1993a, 1997a, 2003a for examples in Old and Middle Scots). This places variation and change in historical Scots in a linguistic vacuum, creating an arbitrary sense of the separateness of various factors influencing historical change. This has also resulted in different predictors being put forward as the key effect constraining or facilitating the anglicisation of Scots. For example, Devitt (1989a) chose to examine only the correlations between genre and change over time, and accordingly suggested that genre was the main factor influencing anglicisation. Meurman-Solin (1989a, 1992, 1993a, 1997a) has examined a large number of extralinguistic predictors, but has done so on an individual basis without taking into account the holistic nature of language change, and the potential collinearity between several of her predictors. Furthermore, “unexplained variation” identified in subjective analyses may actually stem from the multidimensional nature of textual registers - this variation may be subject to other constraints that are not accessible in a single-factor analysis (Nevalainen, 2006: 566).

A significant development was made with Romaine’s (1982) analysis of <wh-> deletion in relatives in Middle Scots, which utilised the previously untapped potential of regression modelling in diachronic analyses. By using the Variable Rule programme VARBRUL (Sankoff, 1975), Romaine was able to observe the effects of multiple (rather than individual) predictors on <wh-> relative deletion. This

represented a valuable move forward in historical Scots research. However, VARBRUL is unable to include author and text as a random effect. Random effects recognise that there is always ‘random’ variation present in a data set, which can be attributed to the idiolect of particular authors or the particular manifestation of a variant (Johnson, 2009). Without their incorporation the regression model fails to recognise where a possibly crucial source of the variation comes from, and potentially overpredicts the significance of the independent variables (the fixed effects) in the model (Baayen, 2010; Baayen et al., 2008; Johnson, 2009; Tagliamonte & Baayen, 2012). Indeed, Romaine (1982: 207) noted herself that; ‘the multivariate analysis may conceal as much as it can reveal’. Furthermore, since Romaine’s publication, there has been considerable progress made in the application of statistical methods to variable data.

In a previous study (van Eyndhoven & Clark, forthcoming), I examined the change from <quh-> to <wh-> in Middle Scots using data from the Helsinki Corpus of Older Scots (Meurman-Solin, 1995), utilising logistic mixed-effects regression models (with author and word as random effects). This was the first study to incorporate these techniques into research on Historical Scots. We discovered that it was audience, rather than text-type (which has been previously claimed – see Aitken, 1979; Devitt, 1989a; Görlach, 1998; Meurman-Solin, 1989b, 1992, 1993a, 1994) that was primarily driving the change that took place for this variant. This indicated the fresh insights available to the researcher through use of current statistical techniques, and the need for more up-to-date methods for approaching historical Scots.

In terms of research on eighteenth century Scots, both quantitative and statistical analyses are few and far between. Those that have taken a quasi-quantitative approach seem to have been largely based on raw frequencies taken from a very small dataset (Cruickshank, 2012, 2017; Corbett, 2013). It is not yet clear which factors were most influential in driving or determining eighteenth and early nineteenth century Scots usage, as there has been no empirical analysis of a larger cross-section of Scots society during this time period. Yet, the various statistical methods now available have

different strengths that can be used to match particular kinds of data to examine effects on the macro and micro scale. We can both analyse large corpora with multiple authors, but also select individuals and their specific writings. Such tools are much more capable of capturing the multifarious and heterogeneous nature of historical data, the antipodal pressures stemming from local and supra-regional interests, and the fluctuating data that characterises diachronic corpora, creating greater scope to pinpoint possible factors influencing language change at specific moments in time. By adopting current statistical modelling techniques, we can reach a better explanatory account of the socio-historical factors which promoted or inhibited language change in eighteenth century Scots. Accordingly, such methods will be employed here.

The specific benefits of statistical modelling and their application to my data will be discussed further in section five (results and discussion). First however, I will give an overview of the eighteenth century and the events that led to it being such an interesting time period; historically, linguistically and socially (section 2.0). Following this I will present the research questions that arise out of this literature and the ideas they seek to explore more detail. In section four I will outline my methodology and the complex steps involved in building a new corpus and searching its contents. Section five contains first a discussion concerning the use and benefits of statistical methodologies, followed by an explanation of how these operate, along with the results they generated, and ensuing discussion. The significance of the results is explored in section six. This is followed by the limitations of the project and possible avenues for further research in section seven, and concluded with closing remarks in section eight.

2.0 Literature Review

2.1 Historical Background - Linguistic Change

2.1.1 The Union of the Parliaments, 1707

Up until the end of the 16th century Scots was the predominant language of lowland Scotland (while Gaelic was still predominant in the northern Highlands), and was the language of the courts, church, legal proceedings and literature (Meurman-Solin, 1993a; Bugaj, 2004a; Devitt, 1989a; Douglas, 2001; Romaine, 1982; Pollner, 2000). There is clear evidence that it was well on its way to becoming a Standard language (Meurman-Solin, 1993a; Millar, 2005; Görlach, 1998; McArthur, 1979; Johnston, 1997; Bugaj, 2004a, 2005). However, the Union of the Crowns in 1603 united Scotland and England under King James VI, and, as the court moved to London, so Scotland's symbol of power and prestige moved south. The Standardising processes at work were interrupted, and instead Scots became increasingly anglicised. This is not to say Anglicisation had not been underway prior to the Union, indeed anglicised forms had already been introduced as a result of the Reformation of 1560, the introduction of the printing press, the prestige of English literature and contact with English speakers (Aitken, 1979, 1997; Robinson, 1983; Dossena, 2011).

However, after the Union of 1603, Anglicisation occurred more generally and frequently in the speech and writing of the Scots gentry (Aitken, 1997). The literate elite and upwardly mobile classes now experienced increasing social pressure to adopt English as the language of high society. In order to move within the upper echelon of the London gentry, it was both desirable and increasingly necessary for the elite to speak the language of their southern neighbours. Aitken (1984: 92) has identified in the written language of 17th century Scottish gentry's private correspondence a 'rapidly anglicising, mixed kind of speech.' For example, the Older Scots regular *-is* plural inflection of nouns, such as *expensis*, *spoussis* competed alongside the Southern form *-s* during the seventeenth century, and had disappeared from Scots by the beginning of the 18th century (Beal, 1997). By the end of the seventeenth century, Scotticisms were all but gone in the correspondence of some of the

upper gentry, (Aitken, 1979, 1997). Furthermore, the Union expanded the audience for literary texts beyond Scotland, a key consideration for many authors attempting to reach a wide audience and ensure their publication was profitable. Indeed, it appears the target audience served as a main motivator driving the switch to certain English variants (van Eyndhoven & Clark, forthcoming), encouraging anglicisation within multiple literary arenas.

By 1707, Anglicisation had already been ongoing for over a century and was well entrenched in most written work. However, the second Union between England and Scotland; the Union of the Parliaments in 1707, significantly boosted the anglicisation process. This Union formally dissolved Scotland's independent parliament, and with the remaining pillar of power and prestige shifted South (Dossena, 2005), there was now little left to keep the gentry in Scotland. As a result, many moved to London or at least spent considerable time there, where they became well established within the elite circles of London (Dossena, 2002, 2005). This in turn added significant value to the prestige of English by reinforcing the status associated with it (Dossena, 2005, 2011; Cruickshank, 2012). The English tongue was considered imperative for ultimate social advancement in the newly developed nation of Great Britain (Smith, 1970; Aitken, 1979; Jones, 1995; Corbett, 2013; Cruickshank, 2017). The upper classes largely embraced the new concept of Britishness, attempting to make an important cultural statement and utilise the new routes to promotion and position (Smith, 2007; Crawford, 1992; Phillipson, 1970; Davidson, 2003). There were many individuals and groups who truly believed that the political union had brought with it the sense that a national British language would be a major cultural, social and economic gain (Jones, 1995). James Buchanan, a linguistic observer and commentator, suggested furthermore that correct pronunciation would also remove the barriers between Scotland and England that were chiefly fostered through different forms of speech. He believed that a shared language could connect the two nations much more strongly than any political union (Crowley, 1991; Dossena, 2005).

It appears that the link between political change and language use was already identified early on during the eighteenth century, adding impetus to the anglicising trends of the previous century and further increasing the status of English. We could expect most members of the elite therefore to show high levels of anglicisation early on in the piece, although the general frequency of Scots in the writings of the literate has not been adequately explored. Yet these appear to have contained many Scots grammatical and lexical features, as they attracted the criticism of the orthoepists; language commentators that took hold of Scotland's literary scene during what is known as the 'Age of Politeness.'²

2.1.2 The Age of Politeness

Throughout the eighteenth century the linguistic divide deepened as the Union ushered in the Augustinian culture of England, accompanied by the 'Age of Politeness' (Aitken, 1984; Beal, 1997). Scots experienced greater linguistic awareness and self-consciousness towards their own speech (Dossena, 2005) and it became desirable to speak 'Polite English'; a highly codified language necessary for any self-respecting gentleman, especially if he wished to achieve an equal partnership with England (Cruickshank, 2012, 2017; Smith, 1970). Some Scots were openly mocked for their speech both at home and abroad by English speakers, a demoralizing blow for a country already struggling with various socio-political issues (such as the Jacobite uprisings and Highland clearances) and poor economic growth (Phillipson & Mitchison, 1970; Templeton & Aitken, 1973). Furthermore, as Mitchell's (2012) study of 18th century English Grammar books for 'foreigners' has shown, there was a prevalent attitude within England at this time that correct language pronunciation equated to good character and constitution. In order to be accepted, outsiders had to learn the English language

² In this sense we are using 'politeness' to refer to the common term used during the eighteenth century rather than the contemporary linguistic understanding of politeness (see for example, Brown & Levinson, 1987; Watts et al., 1992; Meier, 1997). 'Politeness' during the eighteenth century referred to set of conventions dictating correct behaviour, mannerisms and speech in order to be part of high society. Although such ideas had always been around, they took on new definitions and prominence during the eighteenth century, especially in Scotland which was felt by many to be 'primitive' and 'backwards' compared to their English neighbours.

correctly and display competency in grammar and pronunciation. Those who failed to do so were not exhibiting allegiance, commitment and conformity, and were accordingly excluded from English society (Mitchell, 2012: 123). Non-standard speech, such as Scots, was equated with roughness or the hallmarks of a backwards society (Mitchell, 2012).

As a result, the Scots tongue came under scrutiny and was increasingly rejected by the socially mobile classes in Scotland as sullyng and degrading, associated only with conservatives, eccentrics and the common or 'vulgar' people (Aitken, 1979; Smith, 1970; Millar, 2003). Language usage became split along class rather than regional-based lines, and the gap between prestige and context-suitability became significantly wider through the combined effort of the Union and the Age of Politeness (Dossena, 2012). Such attitudes have been observed time and again in sociolinguistic studies; speakers accommodate their speech style as a means of acquiring social approval and maintaining a positive social identity. This accommodation will be in line with the speech characteristics of the interlocutors (Llamas et al., 2009). In this case, Scotsmen were attempting to accommodate to Southern English models, which was not only spoken by the upper English classes they sought to join, but was simultaneously being upheld as the only language style fit for their social class.

This rapid push towards speaking standard English was supported by a network of educational and social societies. Within Edinburgh, social life was largely enjoyed in literary clubs and these became an integral feature of the gentry's social education. Many of these focused on improving manners, literature and conversation, modelling their meetings on those held in London coffee houses. They embraced English literature, newspapers (Harris, 2005a) and the Augustinian values of refinement and propriety. Various societies and language schools such as the *Select Society* sprung up, intent on educating the literary masses on correct pronunciation, grammar, vocabulary and nuance whilst eradicating the 'barbaric relic[s] of a backwards society' (Jones, 1995, 1). They argued that the Scots tongue only served to impose an unfavourable barrier to social success:

‘As the intercourse between this part of GREAT-BRITAIN and the Capital daily increases, both on account of business and amusement, and must still go on increasing, gentlemen educated in SCOTLAND have long been sensible of the disadvantages under which they labour, from their imperfect knowledge of the ENGLISH TONGUE, and the impropriety with which they speak it.’

Select Society of Edinburgh, 1761.

The eighteenth century saw various attempts to correct this situation, including elocution lessons, spelling books, guides and printed lists of Scotticisms produced by notable Scotsmen such as David Hume (1711-1776), James Beattie (1735-1803) and John Sinclair (1754-1835), who sought to remove all traces of their Scottish origins in their speech and writing (Aitken, 1979; Murison, 1979; Smith, 1970; Caie, 2007). This obsessive attitude was present in the linguistic consciousness of some of the Scottish intelligentsia well before the Augustinian Age (Aitken, 1979), and it appears that many of the elite began to feel that the Scots tongue was more suited to homely, domestic spheres of use, rather than any form of dignified writing, already early on in the piece (Jones, 1995; Aitken, 1979). However, by 1755 it had become a linguistic ‘witch-hunt’ (Dossena, 2005: 74). Hume and Beattie often corrected each other’s works, or sent drafts to be proof-read before publication, and frequently critiqued works produced by their contemporaries (Templeton & Aitken, 1973; Aitken, 1979). This may have created a publication standard for emerging Scottish writers and thus paved the way for such practices to continue (Dossena, 2011).

Furthermore, the Scottish public, at one million people, was not sufficiently large for the support of professional authors (Graham, 1908; Clive, 1970) and so the choice often became binary; either authors wrote in English and published to English audiences, or they engaged in other professions alongside. As a result, some of the most outstanding literary works of the eighteenth century were produced by lawyers, clergymen and professors (Clive, 1970; Craig, 1961), but many turned to

English models instead (Dossena, 2002, 2011). Scottish intellectuals sought recognition on a nationwide basis, and this relied on English models of work (Dossena, 2012).

Alongside this puritanical linguistic movement, the eighteenth century saw the birth of the Scottish Enlightenment, a period of remarkable intellectual thought and discovery that saw the focus of academic thought shift from closed national borders to the much wider, European audience (Dossena, 2005). Yet this golden era was also notable for its paradoxical nature; despite its profound literary and intellectual achievements, Scottish society was plagued by a deep insecurity and confusion towards its own language and identity, coupled with a persistent sense of inferiority (Clive, 1970; McClure, 1994: 40; Bono, 1989; Dossena, 2002; 2011). Accordingly, English effectively became the official language for all literary spheres and serious writers, and Scots became increasingly restricted in use and scope, no longer used as medium of everyday writing (Murison, 1979; Corbett, 2013).

Such a picture would seem to suggest a steep decline in Scots and the eventual abandonment of the language altogether. The political and social benefit of adopting the English standard was clearly pertinent to the considerations facing the elite; the fields of politics and language were closely connected and intertwined throughout this time. As the Union brought with it increased mobility, it was not enough to simply relocate to the prestigious South, social mobility required the adoption of the English standard. In light of such pressures the effect of the Union agreement could be expected to be that of Scots language death. Yet this was not the case. Some members of elite must have behaved differently, or have been influenced by other factors that caused them to retain low levels of Scots in their writings. Though it is as yet unclear whether political affiliation or an underlying sense of patriotism played any role, a number of factors have been identified in previous research as responsible for the survival of Scots during this time period, including the emergence of Scottish Standard English, the rise in antiquarianism and the vernacular revival.

2.1.3 Scottish Standard English

Not all voices were uniform in denouncing Scots, and some despised the practice as unpatriotic and insulting to the historical integrity of the language. There were plenty of concerned commentators arguing for the preservation of Scots, whilst the Enlightenment also stimulated discussion regarding language and identity (Jones, 1995, 1997; Millar, 2013). Many who disliked the anglicizing onslaught were aware of the advantages in adopting a London metropolitan standard, but also realized there was a strong practical case to be made for the preservation of Scots (Jones, 1995). The end of the eighteenth century saw the obsessive nature of dialect suppression weaken with a new wave of Scots romantic writers and a rise in antiquarianism and Scots patriotism (Graham, 1908; Aitken, 1979). One of the most important publications arising out of this time was Rev Dr John Jamieson's (1759 – 1838) *Etymological Dictionary of the Scots Language* (1808) (Dossena, 2002). Jamieson denounced the linguistic normalisers as excessive perfectionists and produced a lengthy dissertation on the origin of the Scots language, to assert its equal status with Standard English.

Furthermore, towards the end of the century it appears there was increasing recognition and confidence in a developing Scottish standard emerging out of the language of the legal, educated and clerical circles of Edinburgh, that was equally acceptable to Scottish polite society (Aitken, 1979; Dossena, 2011; Smith, 1996, 2007; Corbett, 2013; Jones, 1995). Aitken (1997) suggests Scottish Standard English first emerged near the end of the seventeenth century and this developed during the eighteenth century, attracting the attention of contemporaries such as Adam Smith and James Adams (Jones, 1991, 1993, 1995 & 1997a), who considered the 'tempered medium' (Adams, 1799: 157) equally appropriate for the educated and professional classes. There was thus no need to import a metropolitan standard since there was already a prestigious native form flourishing (Jones, 1995).

This standard maintained various lexical items as a result of the independent Scottish Church, local government structure and legal, educational and electoral systems that were protected by the Union

of 1707 (Phillipson & Mitchison, 1970; Murdoch & Young, 2007; McCrone, 2007). The First Statistical Account of Scotland provides evidence of the retention of particular Scots legal and clerical terms throughout the eighteenth century (Millar, 2003), reflecting cultural differences that were integral to a Scottish way of life (Cruickshank, 2012; Dossena, 2005). Bugaj (2005, 2013) and Kopaczyk (2012, 2013) have found that Middle Scots legal documents maintained Scots forms much longer than other prose material of the same time period, as they contained codified expressions and specialised lexis which were necessary to construct legally valid texts. This as such allowed certain Scotticisms to continue in the speech of the elite, who were often concerned with buying land, goods and services and who effectively funded the local parish church (Cruickshank, 2012).

Cruickshank's (2012) analysis of the Fife-Rose corpus indicated that Scots was often indispensable, even to the elite. Cruickshank examined the correspondence between James Duff, the 2nd Earl of Fife and a wealthy land owner, to his factor (a trader who receives and sells goods on commission) William Rose between the years 1764-1789. Despite operating within high society and spending considerable time outside of Scotland, it appears Fife relied on particular Scotticisms to provide clarity to a request, particularly when this concerned matters closely tied with Scottish institutions or a Scottish way of life. Scots lexical items could also provide particular pragmatic or semantic meaning. Fife used Scotticisms when a demand was being made or the recipient was being scolded, as these could soften the imposition or suggest a certain familiarity or friendliness. Fife also utilised the semantic extension that Scots could give to his speech, using Scots words to make a semantic distinction from the English equivalent, such as *Kirk* and *Church*. Thus, despite the pressures of anglicisation and his social class, Fife's business interests and involvement with the English elite did not prevent him from using a number of lexical items that were integral to a Scottish way of life. This in turn may have offered such words a similar prestige to Standard English, paving the way for Scottish Standard English to emerge. Scotland and Scots were often intricately intertwined, and thus any investigation that seeks to identify differences in language use along nationalistic lines, for

instance, must bear in mind the complexities of the situation characterising eighteenth-century Scotland.

2.1.4 Antiquarianism

Alongside the increasing recognition of Scottish Standard English, the late eighteenth century saw a wave of Antiquarianism that rode on the crest of the excessive anglicising from earlier on (Dossena, 2012). However, this could manifest as a patriotic backlash to the *Select Societies* of earlier, but could also be a direct product of them, by pushing for language preservation rather than continuation. There was certainly a large group who fundamentally disagreed with the contemporary language attitude and sought to reclaim a lost sense of pride, actively advocating Scots and promoting its survival (Jones, 1995). They created the *Society of Antiquaries of Scotland* and there were a number of enthusiastic followers of this position, such as Sylvester Douglas (1743 – 1823), James Elphinston (1721 – 1809), John Callander (1722–1789), Henry Mackenzie (1745 – 1831), Alexander Geddes (1737 – 1802) and James Adams (1737 – 1802) (Jones, 1995). These individuals, far from being apologetic for the apparent ‘impropriety’ or ‘impurity’ in Scots, emphasised the long and prestigious history behind the language and its pristine state of originality (Jones, 1992, 1995: 5; Dossena, 2011). Geddes and Adams argued that Scots had more integrity than English because it preserved its original Saxon better than English (Jones, 1995; Dossena, 2005; 2011) and Adams called for a separate and identifiable Scottish system of orthographic representation, recognizing the threat posed by adopting the southern metropolitan standard. Elphinston and Geddes, contrary to the likes of Hume and Sinclair, sought to reform orthography in their works (Jones, 1995). Adams’ publication *The Pronunciation of the English Language Vindicated* (1799) was a particularly powerful defence of the Scots language and the eloquence and dignity of those who speak it well (Dossena, 2005; Beal, 1997). These arguments became in fact crucial to the defence of the dialect (Smith, 2007; Dossena, 2011), as antiquity and independence became the cornerstones on which linguistic respectability was built (Dossena, 2012).

Antiquarianism and the reassessment of original linguistic features became increasingly associated with patriotism and sentimentality (Jones, 1997), as political sentiments lent an idealised overtone to Scots (Dossena, 2005). MacDonald (2011) has suggested that the republishing of many Scottish works from the Middle Ages was not simply an act of cultural nostalgia, but rather a reaffirmation of Scotland's own identity following the Union, in consequence of the controversies resulting from the event. Antiquarianism became connected with nationalism, and the search for Scotland's historical integrity implicitly meant a search for authenticity, originality and status as if 'almost to compensate for the disappointment that the Union had engendered' (Dossena, 2005: 129). Scottish literary culture found renewed vigour by rediscovering its roots, manifesting in an increased interest in and publication of the ancient classics, a fascination with Scottish historiography and the links between the Scots language and 'Scottish virtues', and the 'rediscovery' of the vernacular literature and songs of Scotland (Dossena, 2005; MacDonald, 2011). Millar's (2004) research into the First Statistical Account of Scotland found that use of local words often situated the discussion within an existent or desired history, suggesting a certain romanticising of a heroic, glorious past.

Scots and national identity became more closely connected, and political and national leanings interacted with linguistic attitudes on new levels, reflecting general cultural changes taking place during the eighteenth century (Dossena, 2005). This would suggest a conscious connection between political affiliation and language use. Those who displayed nationalistic and patriotic sentiments, who disagreed with the Union and who wished to return to independence were also often those who sought to preserve and maintain Scots, and who disagreed with the anglicising efforts of the orthoepists. Given the emerging association between language and nationalism, a quantitative difference in language use between authors from either side of the political divide is conceivable, especially in light of the antithetical political and linguistic positions characterising eighteenth century individuals.

Indeed, this antithetical stance is apparent in the nature of the other Antiquarian camp, which did not take a supportive role for Scots, but rather saw it as a language to fossilize (Aitken, 1990). The majority establishment position in Scotland from the end of the eighteenth century suffered from *Pinkerton Syndrome*. John Pinkerton (1758 – 1826) wanted to preserve Scots as an ancient and poetic language but not a living one, and was the editor of *Ancient Scottish Poems*; a selection of Older Scots poems carefully collected by the scholar (Aitken, 1990). Indeed, alongside the independent ecclesiastical, legal and educational institutions of Scotland; poetry was the only other literary arena where Scots features were accepted by polite society (MacDonald, 2011). Their interest stemmed from a historical, though also patriotic, perspective, and they had little interest in the Scots of their contemporaries (Millar, 2013). This group of grammarians and linguistic commentators were seeking nothing short of a language death situation (Jones, 1995), and focussed primarily on the republication of earlier works, with the hope to preserve a language long since passed. To some extent they appear to have been successful; the First Statistical Account of Scotland indicates that Scots tends to be associated with the historic and the quaint, the lowly, rural and rustic (Robinson, 1972; Millar, 2012), whilst sometimes Scots could be used to give a little ‘local colour’ to literature (Millar, 2004, 2013; 322). The divide between the ‘rustic’, rural and traditional dialect forms and what was increasingly seen as the coarse, urban working-class dialects became more marked during this time (Dossena, 2005; Millar, 2013).

Yet the sense of inferiority so keenly felt in the beginnings of the century was largely replaced with the increased antiquarian interest in ancient lore, proverbs and traditions, as well as an emergent popular culture based around a historic, romantic vision of Scotland and the Highlands (Dossena, 2005). This was crucial in maintaining some kind of status for Scots, which was able to tap into new vitality and acceptability through the medium of creative literature (Smith, 1970). It appears as a literary medium Scots was accepted in certain genres by the dominant educated opinion, but not as a medium of every day formal conversation (Dossena, 2012). This did however, allow Scots to

develop, bloom and resist the pressures of anglicisation in the realm of creative literature, to create what is known as the ‘vernacular revival’ or ‘backlash.’

2.1.5 Vernacular Revival

Alongside the patriotic sentiment that arose to challenge the anglicising zealots of the eighteenth century, linguistic resistance also found a strong voice through the medium of poetry. The ‘vernacular revival’ of the eighteenth century and the works of some of the greats of Scottish poetry – Allan Ramsay, Robert Fergusson and Robert Burns - have frequently been depicted as the hallmarks of a patriotic backlash to the anglicising tide (Murison, 1979; Aitken, 1979; Smith, 1970). Yet to label this period a ‘vernacular revival’ is somewhat of a misnomer, given that Scots had persisted as a literary medium prior to this flourishing of literature (Robinson, 1973; Beal, 1997). Rather than a “revival” of Scots, it became a reacquisition of some of the status it had lost through the prescriptivism from earlier, gaining covert prestige while losing overt prestige (Dossena, 2002; 2005). Regardless, it is clear that the eighteenth century saw one of the most impressive periods of Scottish literature, a flourishing that has been unparalleled ever since. Poets, in particular Ramsay and Burns, shed new light on Scots and restored its sense of dignity as a contemporary literary language. They helped to create an extraordinarily popular vernacular literature and a market for such works (Dossena, 2005). Ramsay’s antiquarian attraction to Scots, as well as admiration for the expression possible through his own language, led him to distance his poetry from the anglicized, polite and classifying trends of the century (MacDonald, 2011).

However, the motives of the poets were largely creative rather than nationalistic (Jack, 1997). Whilst there was frequently an element of patriotism in the gesture, it was often as much a case of poetic necessity as sentiment. Poets relying on imaginative language sought words that stemmed from a lifetime of experience, tradition and feeling based in Scotland, and often no other word would do (Craig, 1961). Furthermore, poets were innovative with their linguistic repertoire, incorporating both English and Scots graphemes and lexis to broaden their creative range, creating a kind of Anglo-

Scottish hybrid (Buffoni, 1992: 127; Dossena, 2005: 96; Corbett, 2013). McClure (1987, 1996) has pointed out that it would be too simplistic to see poets' linguistic choices in terms of a binary distinction between Scots and English; many works exhibited the full continuum from English to Scots with Anglicised Scots in between. Corbett's (2013) study of the spelling practices of Ramsay and Burns indicated the use of innovative features; in some cases English graphemes were used to reduce the unfamiliarity of Scots words, or to indicate a Scots pronunciation when the item is shared between Scots and English. It seems poets were developing and refashioning Modern Scots orthography as a system in its own right and drew on a range of literary and linguistic resources to do so (Corbett, 2013).

Furthermore, poets may have adopted different linguistic resources for particular registers and contexts to create a certain effect (Beal, 1997). Burns became extremely skilled in moving across the continuum to achieve different stylistic effects, equating Scots with personal and local experience and English with more general ideas. He thus associated meaning with choice of vocabulary, in accordance with the social situation of his language (Craig, 1961; Smith, 1970; Dossena, 2005, 2012; Smith, 1996, 2007). Ramsay also made use of the continuum for different registers and to express particular themes, preferring Scots for satire and farce, and Fergusson similarly utilised the relationship between the broad and polite to make a rhetorical point and broaden the creative boundaries of his work (Corbett, 2013).

Their works suggest a creative repositioning of the different language varieties, although this also reinforced the idea that Scots was only suitable for imaginative writing (Dossena, 2005). This practice had already been occurring in earlier Scottish literature and it appears many eighteenth century literati were skilled in dialect switching and style drifting, but this took on new and increased vitality during this time, particularly as Scottish literature expanded outside Scotland's borders (Aitken, 1979; Smith, 1996; Corbett, 2013). Poets were constrained by intelligibility also; texts that were too Scots-heavy were simply out of reach to the English-speaking populace, unless they were

accompanied by an extensive glossary. Indeed, the English poet William Cowper wrote to Samuel Rose claiming that he hoped Burns would discard his ‘uncouth dialect’ in his poems (Dossena, 2005: 99). This may have led to the use of English for patriotic sentiment, in order to enlighten an English audience of the roots of such patriotism (McClure, 1987, 1996).

Nevertheless, there is no denying the perceptible link between broad political changes, language normalisation and the flourishing of native poetry. The mounting dissatisfaction with the Union agreement and the increased anglicisation that followed in its wake, created an intensified patriotism and political awakening that was often best expressed through cultural outlets (Craig, 1961; Clive, 1970). Many authors utilised the medium of poetry as a covert social commentary on political affairs or to make veiled patriotic remarks (Dossena, 2005; Smith, 2007). Indeed, Robinson (1973) has argued that the revival of Scots poetry during the eighteenth century was in fact largely caused by the Union of the Parliaments. As Scotland was finally stripped of a separate identity, there was a simultaneous backlash of patriotic nostalgia which found an outlet in antiquarianism. The poems of Burns and contemporaries became icons that were simultaneously emblems of patriotism and sentimentality (Dossena, 2012). Ramsay had a rooted nationalism and mourned the loss of Scotland’s political independence but was determined her poetry wouldn’t follow. He sought to highlight the rich history and considerable weight of Scottish works through his collection of Scottish proverbs and songs from the Middle Ages and Renaissance (MacDonald, 2011), whilst maintaining Scotland’s literature through the production of his Scots poems (Smith, 1970).

Poets could be both defensive and assertive in their language use, and Fergusson’s patriotic ideas were often explicitly mentioned in his works:

‘Black be the day that e’er to England’s ground

Scotland was ekit by the UNION’s bond’

(The Ghaists: A Kirk-yard Eclogue)

Fergusson and others felt the Union was less than advantageous for Scotland. Some regretted the Jacobite defeat and many idealised a romantic, independent past that had been lost to the Union (Gibbs, 2006; Dossena, 2012). Their works inspired later generations of poets with similar sentiments, not least that of Robert Burns, who openly supported the French revolution and compared it to the Scots' victory at Bannockburn in 1314. Some poets sought to distance themselves from the overtly English models through expressive use of Scots, utilising positive in-group identity markers (Llamas et al., 2009) to signal their allegiance to Scots and Scotland.

Yet it seems that the linguistic choices of poets and songwriters were sanctioned by the elite, as long as their opinions were directed into specific literary channels. A fundamental shift in the boundaries of acceptable language in written domains occurred during this time (Dossena, 2002, 2012; Millar, 2013), and thus textual mediums such as poetry became valuable arenas to air political grievances but were also one of the few places where this was tolerated by the establishment. This interplay between resistance to the established order (both linguistic and political), and the simultaneous toleration, if not acceptance, by the established order, adds to the complexity of the eighteenth century. Again, as with the antiquarians, the use of Scots and patriotism are closely aligned. Within the field of creative literature, this is expressed perhaps even more covertly than within the realm of serious antiquarian prose. Although there have been various studies examining the works and the motives of Ramsay and Burns in particular (see Clive, 1970; Robinson, 1973; McClure, 1980; Beal, 1997; Jones, 1997; Corbett et al., 2003; Dossena, 2012; Corbett, 2013), there has been less focus upon a wider range creative works being produced during this time.

Thus, it is unclear whether the contemporaries of Burns expressed the same degree, if not proficiency, of Scots in their works. Nor is it yet clear whether the expression of patriotism through poetry and such-like encouraged the use of Scots, and whether this translates into observable,

quantificational differences when comparing works with a political, nationalistic focus, to non-political, creative works.

But, for such an association to exist, there must have existed varying political sentiments, which formed in reaction to major political changes occurring throughout the century. The eighteenth century was characterised by changeable, divergent and often turbulent political changes, and these are discussed in more detail below.

2.2 Historical Background - Political Change

2.2.1 Political Tension

Adding to the complex and the increasing linguistic resistance, was a gradual build-up of political tension occurring during the eighteenth century, aided by increased public participation in political affairs and a growing political awareness among the nation (Hutchinson, 2017). Despite the opportunities for advancement and trade that came with the Union (Gibbs, 2006), relations with England were not always smooth. Initially the Union was welcomed by many and some, such as Sir Walter Scott, believed the Union would heal the divides caused by the Highland/Lowland³ split through their incorporation into a new, unified nation. Both groups could contribute to a common cause which would finally remove the entrenched separation between them (Gibbs, 2006).

However, the eighteenth century saw the eruption of the Jacobite Risings which rejected the Union and further polarised the split between the two groups, entrenching certain hostilities. The Jacobite Risings of 1708, 1715, 1719 and 1745 were based predominantly in the Highlands, with the aim of returning James II of England and VII of Scotland, and later his descendants of the House of Stuart, to the throne. Although the risings have often been portrayed as a strictly Highland phenomenon,

³ The Highland/Lowland split or divide refers to a historical division within Scotland. This was geographical but also social; the Highlands maintained the clan structure of social organisation and continued to be Gaelic speaking well into the eighteenth century. The Lowlands on the other hand became industrialised earlier on, and largely replaced Gaelic with Lowland Scots by the 16th century.

there is evidence that significant numbers of Lowlanders were involved for reasons of their own⁴ (Davidson, 2003). For instance, many members of the landed classes assumed the restoration would reverse or stabilise the effects of the Union, which frequently became the scapegoat for the economic stagnation and reduced power structure they faced in the decades following the treaty. Yet despite involving a large cross-section of Scottish society, the rebellions were perceived by most Lowlanders as product of the volatile Highlands, strengthening their distrust and animosity towards the Highlanders.

Furthermore, despite its promises, the immediate effect of the Union was increased taxation and loss of French trade (Clive, 1970). Scotland was already struggling economically by the time of the Union, and the beginnings of the eighteenth century saw frequent unrest and instability (Phillipson & Mitchison, 1970; Whatley, 2000). There were a wide variety of reactions to the event itself (Murdoch, 2008), with frequent turmoil under the surface that occasionally erupted into open displays of opposition, such as the Shawfield and Porteous riots of 1725 and 1736 respectively (Phillipson & Mitchison, 1970) and anti-English riots before and after the Union (Clive, 1970). Certainly, the Jacobite Risings are a testimony to turbulent socio-political times. Although Scotland was more peaceful than the decades of the seventeenth century and economic prosperity increased after 1750 at a rapid rate (Whatley, 2000; Gibbs, 2006), the memory of the earlier unrest remained (Clive 1970).

The Union had been intended as a partnership, yet the relationship between the two nations often seemed difficult and uneasy, and there was a pervasive sense of unfairness (Smith, 1970). Many felt that Scotland was being denied access to the benefits supposed to be conferred under the Union, and that she enjoyed no popular representation within Westminster, leaving her demands largely

⁴ Davidson (2003) suggests the Lowlanders involved had varying motivations depending on their social standing, including a popular desire to defend Scottish liberties from arbitrary power, the wish to reverse the slow decline of the ruling class in Scotland and the economic blow dealt to Scotland immediately following the Union. For a certain section of the landed classes, the rebellions were seen as a vehicle to express a Scottish national identity that was denied institutional expression as a result of the Union. (Pittock, 2001; Davidson, 2003).

ignored (Pentland, 2008). There was also a notable inequality in the provisions made for the Church of Scotland compared to the Church of England; much less funding was made available to the former than the latter (Smith, 1970; Harris, 2005a). Such discrepancies no doubt fed into the growing disillusion with the Union, whose promises of economic opportunity and benefit seemed increasingly dubious.

The Union did in fact provide many Scotsmen with the opportunity to become involved in political life south of the border, and many did so, often with notable success, but they encountered frequent discrimination and hostility from their southern neighbours, causing widespread resentment and bitterness (Gibbs, 2006; Smith, 1970; Clive, 1970). Anti-Scots antipathy was stirred up by John Wilkes, an Englishman with an entrenched hatred towards the influx of the Scots and their apparent 'takeover' of the English administration, causing protests and rallies across England (Gibbs, 2006). This hostility was heightened during the anti-Bute agitation of the 1760s when Lord Bute, a Scotsman, became the exceedingly unpopular Prime Minister of England (Graham, 1908). Such antipathy amplified the negative view of Scotland that had followed the Jacobite risings (Smith, 1970; Jones, 1995).

This undercurrent of anti-Scots bias continued on into the beginning of the nineteenth century, surfacing occasionally during moments of political turmoil, such as the impeachment of Dundas (Hutchison, 2017). Political discussion and contemporary newspapers south of the border tactically appealed to this long-established popular prejudice within the English nation, when Scottish political affairs were seen to endanger the political balance, as well as feeding the negative stance towards Scotland in general (Hutchison, 2017). Scots felt especially conscious of being Scottish when in London, but their own patriotism was strong and sharpened by English criticism and hostility (Smith, 1970). They often sought out other Scotsmen for company in London, in social clubs such as the British Coffee House (Graham, 1908), and a distinct separation between the two groups remained during the eighteenth century.

Agitation also grew in reaction to breaches in the Union treaty, most notably the 1785 agitation against the Judges Bill (Phillipson, 1970; Bono, 1989). This bill sought to reduce the number of judges on the Court of Session from fifteen down to ten, in order that the remaining judges' wages could be increased without the need to redistribute more funding from the treasury. However, this decreased level of representation triggered widespread hostility and discussions concerning the Union and independence. There was a real fear that if one bill was passed which directly violated the Union, more could follow (Phillipson, 1970). The legal system of Scotland had been left relatively intact following the Union, and this was seen as a direct attack against a fundamental component of Scottish life. Any ministerial attempts to reform national institutions or alter Scotland's political rights were seen as unwelcome and a threat to the Scottish gentry's position as Scotland's governing class. Englishmen were frequently felt to be encroaching on the Scottish political scene, particularly when they sought to alter the regulations surrounding Scottish law, education or religion. Another recurring complaint was the Militia Acts of 1757 (Harris, 2005a), which applied only to England and which suggested the government did not wish to arm the nation responsible for the Jacobite uprising of 1745 (Smith, 1970). Such acts of legislation drew overt comment and lengthy correspondence in the press, with exchanges of views made public (Harris, 2005b).

This frustration, hostility and resentment was present and experienced in various sectors of society throughout Scotland, suggesting these events not only affected large areas of the country, but the concern they generated was also shared. It is conceivable that these shared grievances, with their obvious anti-union aspect, could have translated to the language use of individuals across society. Large sectors of upper Scottish society were becoming both politically and linguistically aware, leading to a heightened awareness of their nation and the language that went with it. It is as yet unclear whether the general, sweeping changes occurring in the wake of the Union were mirrored in the language use of Scottish literate society, or whether this applied perhaps only to particular individuals. An obvious candidate for this effect were the radicals, who emerged with increasing force during the latter half of the century.

2.2.2 Radicalism

During the second half of the eighteenth century the tensions already underfoot were fed by a politically-charged climate arising from the American and French revolutions and Irish Home Rule, which drew agreement and sympathy from many sectors of Scottish society (Craig, 1961; Phillipson, 1970; Bono, 1989). The political consciousness of the Scottish people grew as an interest in political affairs increased (Bono, 1989), and discussion of these external events often came to be grounded within local concerns, which took on greater prominence in light of international events (Plassart, 2014). Initially, the French Revolution was seen as a manifestation of new movements towards European enlightenment and global political and religious liberalisation. It generated admiration from the Scottish literati and press alike. However, as Harris' (2005a) analysis of Scottish newspapers has shown, reactions towards the French revolution changed remarkably following the violence and bloodshed from 1793 onwards.

Nonetheless, the Universalist ideas arising from the revolutions circulated throughout Scotland during the latter half of the eighteenth century, and stimulated Scottish radicalism which emerged with increasing weight and force (Pentland, 2004). Absolute parliamentary sovereignty and Scottish semi-independence became frequent topics of discussion, and between 1792-94 radical agitation reached its peak (Pentland, 2011). Political societies and organisations, such as the Scottish Friends of the People, Zetetic Societies and later the more radical United Scotsmen sprung up, and large numbers of people from a wide range of professions became involved due to the low subscription rates (Bono, 1989). These societies sprung up not just in the urban centres but also smaller villages and rural areas (Bono, 1989). These formed primarily to encourage free discussion regarding politics, representation and rationalism.

Notable figures of the radical or anti-establishment movement emerged from such organisations, and were often outwardly opposed the Anglo-Scottish Union of 1707 such as Andrew Fletcher of Saltoun (1655-1716) (Phillipson, 1970) and the radicals James Callender (1758–1803) and Thomas

Muir (1765-1799). Callender in particular feared the king's influence, seeing the encroachment of English influence as a dangerous corruption of the constitution (Bono, 1989), and was unique in pushing for a Scottish nationalist agenda (Pentland, 2011). Muir became the leader of the radical societies; *The Scottish Friends of the People* and the *United Scotsmen*. He pushed for the formation of associations and societies so that people could petition as a united body rather than as individuals, and defended the rights of people to associate freely for political ends (Pentland, 2016). Muir was eventually transported to Botany Bay, Australia for sedition. His trial, along with the other 'Scottish Martyrs' (the unfair trials of a notable radicals) opened up heightened confrontation between the state and the radicals, who used the opportunity to question government and critique the state (Pentland, 2011).

The end of the eighteenth century saw riots, rallies and demonstrations across Britain at large, and in Perth in 1792 a 'Tree of Liberty' was erected along with cries for an end to monarchy and aristocracy (Honeyman, 2008). There was harsh repression of reform groups during the Napoleonic Wars, and many reform societies wound down, or had to go underground, forming clandestine organisations such as the United Scotsmen (Harris, 2005a; Plassart, 2014). This group eventually attempted to rise against the British government, but troops soon crushed the rebellion (Gibbs, 2006). There was also a concerted effort in the loyalist press to convince the labouring classes of the dangers of joining radical societies, and the dire consequences of subverting the natural political leadership. In turn, it seems an anti-radical stance was prominent among skilled labourers and the elite alike (Harris, 2005a). Yet despite the anti-establishment stance taken at times by certain radical groups, there is also evidence of collaboration and communication between like-minded radical groups across the border. The Scottish Friends of the People made contact with the London-based Whig Association of the Friends of the People and various Scottish groups set about creating communication links to their English counterparts (Harris, 2005b). The printing press became their vehicle for both communication and expression, causing the people of Scotland (and further south of

the border) to be exposed to or have access to propaganda from all sides of the divide, and generating a wealth of political literature.

Although the radicals formed a relatively small sector of Scottish literate society overall, their public activity and persona no doubt influenced the Scottish public, or at least made them aware of the increasing political disharmony and debate around ideas of incorporation and independence. The radicals created debate, opening up not just the grounds for general discussion into the particular situation of Scotland, which brought political ideas into the consciousness of the public, but also generating the medium of political debate itself. The subsequent rise political literature and writing brought political ideas to the forefront for the reader and writer, and it seems feasible that nationalistic ideas could manifest in nationalistic language use. How these two factors may have interacted is as yet unclear. What is clear, is the value of the written word to both radicals and loyalists, whose means of dissemination were ultimately characterised by the cheapest form of public literature; the pamphlet.

2.2.2.1 Political Pamphlets

The Union certainly caused a lively and wide-ranging pamphlet war among radical groups, whilst a noticeable increase occurred during the period of profound instability following the American Revolution, when themes of civil liberty, sovereignty, identity and reform became prominent issues both in parliament and the country at large (Bono, 1989; Harris, 2005a; Pentland, 2011). Political treatises and tracts abounded as the Union, the political structure and parliamentary legislation was increasingly questioned in the latter half of the eighteenth century (Bono, 1989; Harris, 2005a, 2005b). Print became seen as the ultimate vehicle of political discussion and information, accessible to an audience larger than ever before, as a result of increased literacy levels and a readership that was no longer purely Scottish, but thoroughly British (Harris, 2005a, 2005b). Accordingly, it became the site of ideological and political struggle, and a reflection of domestic political developments, although the anti-reform, anti-radical side clearly dominated this field (Harris, 2005a). Part of the

appeal was the ability to manipulate printed works to appeal to many different groups of people within Scottish society, on both sides of the divide. This led to widely varying reports concerning domestic events or legislation, and even reports on the very same trial could differ notably in content and tone, as well as language choice and the material chosen to be included or excluded (Pentland, 2016).

Pamphlets were the medium of choice for writers engaged in political debate and controversies, providing a new and easy means to spread political culture and create public opinion towards political and religious controversies such as the Union (Pentland, 2011; Harris, 2005a). Unlike newspapers which were largely controlled by financial pressures and official hostility, radicals could rely on chapbooks (single page newspaper sheets) and pamphlets to propagate their own agendas (Harris, 2005b; Pentland, 2011). Nonetheless, some newspapers such as the *Edinburgh Gazetteer* chose to publish political writings at a low price, despite the risks this entailed (Bono, 1989; Harris, 2005a). Indeed, in spite of the repression and hostility faced by the radical press campaign, a network of radical publishers, printers and booksellers continued to operate well until the end of the eighteenth century (Harris, 2005a, 2005b). The loyalist campaign on the other hand enjoyed strong official support and financial subsidies in Scotland, allowing for much better representation in press and across the nation (Harris, 2005a). A steady stream of loyalist propaganda circulated through Scotland during the 1790s and ranged in scope from songs and dialect pamphlets to sermons and treatises (Harris, 2005b).

It is feasible that radical propaganda might make use of Scots as a marker of national identity, given that some of the works coming out of the radical press were anti-Unionist, anti-royalist and pushed for Scottish independence. Considering the arising awareness of Scots as unique to Scotland and its people, stimulated largely by the antiquarian and vernacular circles active during this time, such sentiments could be expected to align quite well with the aims of the radicals. Furthermore, their target audience was often local, hence the use of Scots could be beneficial in creating public opinion

towards political affairs. The possible motives of the loyalist propaganda machine are less clear, and could depend on how they attempted to appeal to their target audience, though a preference for English would be expected. It is not yet clear whether there was a quantifiable difference between the two groups and their use of Scots (if Scots was indeed used seriously at all). Nor has there been any empirical study to date which has indicated whether such documents in eighteenth century Scotland contained more or less Scots overall, relative to other genres.

2.2.2.1.1 Scots in Political Pamphlets

Some Scots usage has been identified in both loyalist and radical pamphlets by Pentland (2011). Often, its portrayal could reflect the general division that was slowly arising between Scots and English during the eighteenth century (Pentland, 2011). Thus, anti-radical pamphlets could see the use of Scots in stories concerning the dangers of meddling with politics, describing the degeneration of people into sluggards and deluded workers through Scots verse. This attested the low prestige that was already being ascribed to Scots (Bono, 1989). Radical pamphlets frequently contained dialogue, which was usually rendered in Scots with the aim of effectively communicating and appealing to the popular audience whilst fostering sympathy and a sense of common patriotism (Pentland, 2011). Radical pamphlets also made use of Scots for its communicative appeal and the songs and poems frequently appearing in pamphlet literature had the sanction to go even further in espousing radical sentiments.

Yet, although many radical pamphlets dealt specifically with Scottish issues, the groups, publishers and printers involved in the process were part of a print culture that originated from England and was structured by English models (Harris, 2005a; Pentland, 2011). London became the site of the literary and political culture of Scotland, and Scottish political groups regularly utilised English as well as Scottish newspapers for their arguments (Harris, 2005b). Loyalist literature especially, but also political debate and publications in general, were frequently issued from London. Indeed, several

famous reformists' works by Scots were published first in London, and then the rest of Britain (Harris, 2005b).

The language used within these works was thus liable to intense anglicising pressures, despite their origin, and the use of dialects in such contexts was therefore marked and inevitably self-conscious, considering the default language of print was English (Millar, 2013). It was only later in the period, as shown by Donaldson's (1989) research into Scottish prose during the Victorian period, that newspapers were able to witness a blossoming of new speech-based forms of prose. This was the result of considerable expansion in literary value and the rise of a new, mass literary market within Scotland (Donaldson, 1989). Such freedom was not available to printers and authors of the late eighteenth century, and newspapers in particular were highly anglicised as a result. It remains to be explored whether the works issuing from small radical publishing houses and printers, as well as the language of pamphlets, was anglicised to the same extent, or whether the political goals of the radicals coincided with higher levels of Scots.

2.3 Divergence and Convergence

However, it is easy to posit contemporary structures over historical unrest when in fact the clear, nationalistic divisions we identify today may not have existed in the minds of historical actors. Many of the radicals pushing for reform did not necessarily agitate for a clean break with England, and in fact most sought to legitimise their calls for reform by appealing to an alternative British patriotism (Pentland, 2004). There could be different kinds and degrees of change wished for by reformers (Bono, 1989). Certainly, some were more extreme than others, but it seems the majority of the commentators were divided along the same political lines as their English counterparts (Plassart, 2014). The ideology of the reform movement was anti-parliamentary rather than necessarily anti-English, and there was a high level of interaction between reformers across England, Scotland and

Ireland (Pentland, 2004). Wars with the French reinforced the new idea of Britishness and the notion of 'us' against the 'other', as Scots fought alongside English on the battlefield (Gibbs, 2006).

Radicals sought to appeal to the broadest audience possible in order to air their grievances and calls for reform, and as such laid claim to a flexible British identity to harbour support and understanding rather than taking a narrow nationalistic or separatist approach which could see them side-lined (Pentland, 2004). These actors were after all, part of a political system that was based off a unified nation rather than individual factions, and it was in the very least tactically more sensible to stick to the language of the constitution (Pentland, 2011, 2016; Plassart, 2014). Both radicals and loyalists sought to present themselves as patriotic fighters, lovers of their country and taking such measures precisely because it would benefit the nation (Pentland, 2008; Morton, 1999; Plassart, 2014). In order to engage in political discussion, radicals had to present their arguments in constitutional terms, although they could pursue this by referring to shared and separate histories (Pentland, 2016). Of course, there were also radicals who ultimately rejected a British identity, and some, such as James Callender and Thomas Muir, still became figureheads of the radical movement. Callender in fact disregarded the whole idea of a constitutionalist debate (Pentland, 2016);

“What ‘our most excellent constitution’ may be in theory, I neither know nor care. In practice, it is altogether a CONSPIRACY OF THE RICH AGAINST THE POOR.”

(Political Progress of Britain, 1795)

The eighteenth century thus appears marked by alternating and conflicting views regarding language, politics and identity; various events contributed to movements embracing both linguistic and cultural uniformity and diversity (Jones, 1995: 1). On the one hand there were those who were concerned with social advancement and success, who identified their future within the unified

nation of Great Britain and who identified Scots with the common people or the 'vulgar' (Aitken, 1979: 93; Jones, 1995). Strenuous attempts to imitate southern English models by the elite could be associated with a Unionist agenda. On the other, there were those who rejected the unpatriotic attempts by their fellow countrymen to eradicate the Scots tongue and who disagreed with a union and nation that did not seem to have Scotland's best interests at heart. Similarly, the loyalties of the lower social orders fluctuated significantly throughout this time period; they could be swayed to support the existing social and political system but equally could be convinced to view it with indifference or passive hostility (Harris, 2005a).

2.3.1 Intertwining Influences

Yet the two forces are not as polar opposite as often depicted, and in fact Smith (1970) has shown that some of the most patriotic and nationalistic Scots were also the most ardent supporters and teachers of a 'correct pronunciation.' This is reflective of the general cultural dualism taking place in eighteenth century Scotland. Many Scots attempted to integrate themselves into the new British nation, and yet retained their sense of being Scots (Gibbs, 2006). Although they laid claim to their rights as 'Britons', it seems that being 'British' was largely reserved for special occasions or to achieve certain means; in their everyday consciousness the Scots and English identified themselves as two separate nations (Smith, 1970; Murdoch & Young, 2007; Dossena, 2012).

Simply equating language with patriotism and supra-local loyalties disguises the complex considerations underlying linguistic choices in eighteenth century Scotland. Many linguistic observers professed to be both patriotic Scotsmen as well as strong supporters of the Union (Jones, 1995) and numerous well-known writers indicated equally contradictory feelings about their native dialect (Dossena, 2011). There arose tendencies both to celebrate and to denigrate Scots during this time by leading commentators (Dossena, 2011), and some grammarians such as Beattie made appeals for tolerance towards linguistic plurality, despite recommending the English tongue as the desirable standard; 'To speak with the English, or with the Scotch, accent, is no more praiseworthy,

or blameable, than to be born in England or Scotland' (1788: 91-92; Jones, 1995). James Boswell (1740 – 1795) is a notable example of a Scot who strove to move upwards into London society and took some pains to temper his strong Scottish accent, yet he was reluctant to accept complete Anglicisation (Dossena, 2005: 62-72). Although he encouraged the dissemination of loyalist literature (Harris, 2005a), Boswell also reminisced about Scottish independence and the lost magnificence stolen from Scotland through the Union. He suggested Scotland's love of independence and liberty must continue to be exercised in the eighteenth century (Smith, 1970).

Similarly, Buchanan's English Pronouncing dictionary took pains to focus on 'British' rather than 'English' usage (Dossena, 2012), and there were various attempts to revive historical traditions whilst attaining 'politeness' and social mobility. Such attempts are evident from the publication of dictionaries and histories of the Scots language alongside pronunciation guides and English language manuals by antiquarians and grammarians alike (Aitken, 1979; Jones, 1995). The gentry thus faced a constant dilemma of navigating between the simplicity and purity of nativism, and the other of cosmopolitan sophistication (Clive, 1970).

Furthermore, the linguistic situation during this time was extremely fluid and not a simple case of replacing a set standard with another, with no overlap or intermingling between them (Dossena, 2005). It must be remembered that Scots and English stem ultimately from the same parent language - Old English. Through radical restructuring under Norse influence, Scots had diverged considerably from English (Johnston, 1997), but the development of Scots is marked by a history of lexical borrowing (MacQueen, 1983). Rather than posit a sharp divide between 'Scots' and 'English', many scholars have suggested there existed a general cline from Southern English to Scots (Frank, 1994; McArthur, 1979; Aitken, 1984; Kniezsa, 1997; Görlach, 1996; Kopaczyk, 2012), containing a large common core (Meurman-Solin, 1993a). Accordingly, the distinction between the two languages was by no means clear, allowing a variety of options to be available to speakers in

Scotland at any one time. Linguistic choices could reflect multiple identities as well as tapping into ideas of novelty or specificity.

2.3.2 The retainment of Scots

Most tellingly, despite the most rigorous efforts of the orthoepists to facilitate the complete adoption of Standard English, Scotticisms, including many of those on the lists of grammarians, continued in the speech and writing of Scots both at home and those thoroughly integrated into London life (Jones, 1995: 3; Templeton & Aitken, 1973; Aitken, 1984; Dossena, 2002; Cruickshank, 2017). Aitken (1979: 96) attributes this to the loss of 'linguistic insecurity' and subsiding of the anglicising movement as the eighteenth century wore on, combined with a lack of contact with native Standard English speakers. Millar (2012) argues that there was in fact a deliberate retention of Scots lexis as a result of the literary movement alongside a relatively unconscious interference of Scots structures.

Cruickshank (2012) has suggested that location was also an important factor in the use of Scotticisms; Lord Fife had more trouble avoiding the use of Scotticisms when in Scotland than when mixing with high society in London. The influence of the interlocutor and social setting was pertinent to his linguistic choices, and it is plausible that many Scots felt no real need to omit Scotticisms altogether when in conversation with their countrymen (Cruickshank, 2012). This is further evident in Cruickshank's (2017) study of Lord Fife's letters to the English Prime Minister George Grenville between 1763-1769. Cruickshank found the rate of Scotticisms in this correspondence to be four times less than those in the Fife-Rose corpus, indicating that Fife took greater pains to Anglicise his writings when conversing with London elite than with Scottish. However, despite his efforts, there is evidence in Fife's writings of hypercorrections and persistent Scotticisms, implying a sense of linguistic insecurity on the part of the lord (219). Furthermore, the choice to use Scots was also influenced by pragmatic, semantic and social requirements, whether to manipulate the audience or access a particular pragmatic meaning unique to a Scottish lifestyle (Cruickshank, 2012). Thus,

despite Fife's linguistic caution, he did employ occasional Scotticisms in his correspondence to Grenville for stylistic purposes, particularly when writing in a more familiar style. This suggested closeness and familiarity, which in turn diminished the need to completely suppress the vernacular, as societal acceptance had already been achieved (Cruickshank, 2017: 228).

Finally, as time wore on it seems increasing numbers of Scots speakers did become aware of the value of Scots as a vehicle of expression, feeling, sentimentality and authenticity (Dossena, 2005). Sometimes this took on an openly patriotic and nationalistic sense, as dialect features became increasingly distinct as markers of personal and group identity (Millar, 2013). Moreover, despite his fame and reputation for usage of the Scots language, it must be remembered that Burns was not alone in his choice to use Scots for written and spoken mediums. There were some who were openly proud to speak Scots and refused to accept a confession of inferiority (such as John Ramsay of Ochtertyre, 1736–1814).

Above all, Scots continued to be the spoken language of most people in Lowland Scotland. Furthermore, Donaldson (1989) has shown that newspaper journalists also wrote in Scots during the nineteenth century, to emphasise their sense of exclusiveness and autonomy. This involved every kind of public discourse, including politics at national and international levels. It is clear that vernacular prose is by no means dead during this time, and it remains to be explored which other forms of prose continued to exhibit Scots throughout the eighteenth century. By quantifying the levels of Scots across literate eighteenth century Scottish society, it can become clear where Scots continued and where it faded, as well as who continued to use the language. If Scots was increasingly associated with both political and linguistic resistance, then it seems feasible that we should see a continuation if not increase in Scots among the historic actors who professed such sentiments.

2.3.3 The complexities of the eighteenth century

It appears the linguistic situation was being shaped by opposing forces that nonetheless were frequently tied together in a complex linguistic, political and social power struggle. Various historical actors exhibited different linguistic strategies across various textual mediums to achieve particular ends, and this careful balance was in constant social negotiation. As a result, the division of functions between English and Scots was strengthened and solidified during the eighteenth century, as Scots became increasingly associated with the covert, traditional and close to home, and English with the overt, prestigious and 'proper'. The eighteenth century, more strongly than the decades that had passed before, saw the sanctioning of appropriate literary channels and linguistic behaviour to display identity, patriotism and both local and supra-local loyalties. What is clear in the attitudes of those rejecting or embracing Scots was its position as a marker of cultural identity, both in positive and negative lights. With such opposing forces at work on the language, and conflicts of interest within the hearts of many Scotsmen, the interaction between the historical and political events of the time and the linguistic choices of the authors has the potential to be very dynamic, complex and multifaceted. Yet, previous research examining either the interaction between political change and historical Scots or quantitatively analysing eighteenth century Scots in general are scarce or virtually non-existent.

2.4 Previous Analyses

It is as yet unclear from previous studies, both those focussing on the eighteenth century (Jones, 1995; Aitken, 1979; Robinson, 1973; Dossena, 1997; Smith, 1996, 2007; Millar, 2013; Corbett, 2013) and earlier (Devitt, 1989a; Meurman-Solin, 1989a, 1989b, 1989c, 1992, 1993a, 1997a, 2000b, 2003a; Romaine, 1982), whether a tangible link would have existed between political turmoil and conscious or subconscious use of Scots. The general trend of continuing Anglicisation during the eighteenth

century has been identified (Jones, 1995) and some studies have argued that written Scots was almost wholly subsumed by English by the mid-18th century (Millar, 2004; Murison, 1979). Yet it is plausible to assume that an increase in national awareness and public dissatisfaction may have affected people's use of Scots, and some have suggested that the eighteenth century saw the development of a hybrid language (Aitken; 1979, 1981, 1984, 1997; Buffoni, 1992; Dossena, 2005; Corbett, 2013). This often utilized the orthographic practices of Standard English, but marked out Scots linguistic choices in a variety of ways, such as occasional 'phonetic' spellings that indicated Scottish pronunciations (Corbett, 2013).

The changing nature of writing in Scotland may not have necessarily seen a point-blank removal of all traces of Scottishness from writing, but rather alterations and manipulations along the way. Sociolinguistic studies of dialect contact have suggested salience to be a major factor influencing accommodation. It is often the salient linguistic features that tend to be adjusted, and this in turn can reveal much about their socio-indexicality (Trudgill, 1986). It is not unrealistic to assume that such patterns may be observed in historical data also. Indeed, there is evidence that Scotticisms that fell below the level of consciousness persisted in the writing of the literate, despite the efforts of 18th century Anglicisers (Aitken, 1979, 1984; Jones, 1995; Templeton & Aitken, 1973; Cruickshank, 2012). Cruickshank's (2017) analysis of Fife's correspondence to the English Prime Minister indicated the suppression of the more salient features of the Scottish lexicon, as well as the continuing morphosyntactic influence of Scottish on English. Studies that tend to focus on one or a few salient Scots features therefore have the potential to miss a large amount of 'Scottishness' in writings. Yet, by statistically examining the overall frequencies of a large range of eighteenth-century Scots words, there is potential to discover much higher levels of Scots in writings than previously recognised.

Furthermore, the extent to which the Scots language was associated with patriotism or seen as a vehicle for nationalism by its speakers, is unclear. It has previously been suggested that there was a lack of any clear linguistic loyalty to the Scots language from the majority of people (Jones, 1995;

Aitken, 1979). Certainly, there were those who objected to the Anglicising process, but such indications of resistance do not necessarily tell us much about the influence of political unrest on language choice, as opposed to concerns of a primarily linguistic and stylistic nature. We know that speakers can adjust their use of certain, marked variants to bring themselves closer to their interlocutors, or diverge from them to demonstrate 'social psychological distance' and mark out their affinity with an alternative group (Llamas et al., 2009). It is not yet clear whether the same held for authors and literate Scotsmen writing in the eighteenth century and addressing both local and British audiences, but the possibility is certainly tangible and provides an interesting avenue to explore.

3.0 Research Questions

Despite the incredibly diverse and heterogeneous nature of the eighteenth century, how Scots was affected by these events is still largely unknown. The overall frequency of Scots across society in general during the eighteenth century, and how this relates to the centuries either side is unknown, nor is it clear which factors had the greatest effect on the usage of Scots, and whether this differed for those interacting with the large-scale political changes taking place. Accordingly, the following research questions were formed for this investigation.

1. How did the frequency of Scots lexis pattern over time for the general literate population during the eighteenth century?
2. How did the frequency of Scots lexis pattern over time for politically-active individuals?
3. Which sociolinguistic factors were most important in influencing the frequency of Scots lexis in general society?
4. Which sociolinguistic factors were most important in influencing the frequency of Scots lexis among politically active individuals? Did these differ from that of the general population?
5. Is there an observable difference in usage between political individuals from either side of the Unionist divide? Specifically; did authors who were opposed to the Union use more Scots lexemes than those who supported it?

4.0. Methodology

4.1 The Corpus

4.1.1 Corpus Compilation

This project seeks to analyse the effect of political change on written Scots, and thus requires the writings of particular, politically-involved individuals, as well as texts that reflect political discussion or leanings. Initially, I sought to locate an existing repository of political texts for this research. However, this proved to be impossible to find. Although there are various online Scottish corpora available, both historical and linguistic, none of these in particular have a focus on political works or the writings of eighteenth-century politicians. Even the correspondence of political individuals known to have been active during the time frame in question are scattered across various sources. There are certainly promising collections being developed, such as *The People's Voice* project based at the University of Glasgow (<http://thepeoplesvoice.glasgow.ac.uk/project-team/>), which seeks to create a searchable database of political poetry and songs from the nineteenth and twentieth centuries. However, this had not yet been launched at the time of research, and furthermore was restricted in genre and time period. Similarly, the availability of digitally-converted newspapers, chapbooks and broadsides is problematic, as these are difficult to obtain in large quantities and are limited in their scope. An examination of the online collections held by the National Library of Scotland (www.digital.nls.uk), the Scottish Chapbooks Project at the University of Guelph (scottishchapbooks.lib.uoguelph.ca), the University of Glasgow Special Collections (www.special.lib.gla.ac.uk) and the digitised collection held by the Bodleian Library at Oxford (www.bodley.ox.ac.uk) indicated few papers and documents that dealt with political matter.

Furthermore, in order to ascertain whether political individuals did behave differently, the writings of other, non-political members of the literate Scottish public in the eighteenth and early nineteenth centuries must also be analysed. This creates not only a baseline to enable comparisons between political and non-political language use, but also provides us with a broader, more general

understanding of language use in Scotland during the time in question. This enables us to identify whether this time period was significantly different to the decades on either side of it. If this appears to be the case, then it is plausible that the various socio-political influences operating on late eighteenth-century Scottish society did have an impact overall, at least for the literate sector of Scotland.

Due to the lack of readily available, suitable material, I chose instead to create my own corpus of eighteenth and early nineteenth century political and non-political Scots texts. I firstly required a range of eighteenth-century Scottish texts that covered various genres and authors. Thankfully this could be fulfilled by a pre-existing collection of texts: The Corpus of Modern Scottish Writing (Smith & Corbett, <https://www.scottishcorpus.ac.uk/cmsw/>). This online, freely-available text corpus provided the broad-based, non-political component to my corpus, which shall be referred to henceforth as POLITECS – Political Opposition, Loyalty and Indifference in Texts in Eighteenth Century Scotland. The rest of POLITECS was made up of political texts sourced from various locations. These two components are explained in more detail below.

4.1.1.1 The Corpus of Modern Scottish Writing

The Corpus of Modern Scottish Writing (or CMSW) created by Smith and Corbett (University of Glasgow), is a freely available electronic corpus of approximately 358 documents and 5.5 million words of running text. It spans the years 1700-1945 and covers a range of written and printed texts; including novels, correspondence, newspapers, magazine articles and legal documents such as wills and sasines. Alongside this, the corpus also contains a number of texts produced by orthoepists (language commentators) during the eighteenth century. These figures sought to eradicate the Scots language to ‘improve’ the speech and writing of their contemporaries. Their texts are included in the corpus in order for researchers to compare the orthoepists’ pronunciation guides and recommendations concerning Scots and English with actual language usage. The texts within the

corpus have been divided into 50-year time periods to create five categories in total: 1700-1750, 1750-1800, 1800-1850, 1850-1900 and 1900-1950. This corpus compliments the Helsinki Corpus of Older Scottish Texts (Meurman-Solin, 1995) which covers the time period 1400-1700. The CMSW was available as a series of text files, which were downloaded from the website (<https://www.scottishcorpus.ac.uk/cmsw/search/>). The extra-linguistic information for each document was not always included within the text itself, and so this information had to be requested from the corpus compilers. Wendy Anderson (p.c) kindly sent us the master CSV file, which included information on both the texts (such as genre, publisher, place of publication, and year of publication) and the authors (including their education, place of birth and occupation).

With the exception of the temporal analysis (see section 5.2 below), I chose to look only at texts spanning the years 1700-1860, as that is the time period under investigation. Although this research seeks to analyse whether this particular period behaved differently overall, it also seeks to investigate which extralinguistic factors were most important in predicting the use of Scots during the eighteenth and early nineteenth centuries (e.g. political affiliation, genre, birthplace). Thus, for the purposes of the sociolinguistic investigation, only a subset of this pre-existing corpus was required for the data frame. Accordingly, all texts published after 1860 were excluded. This left 273 texts and 2, 130, 370 words of running text.

4.1.1.2 POLITECS - Political Opposition, Loyalty and Indifference in Texts in Eighteenth Century

Scotland

The CMSW section of POLITECS represents what the literate sector of Scottish society was doing in the wake of the Union, but in order to discern whether there is an effect of political affiliation on language use, texts with a decidedly political focus or background were required. Thus, 29 political documents were sourced from various holdings and added to POLITECS for linguistic analysis. The documents chosen were selected on the basis of the political background of the authors, with a particular focus on those who demonstrated known support or antipathy towards the Union of

1707. The availability of writings by politically-active individuals and politicians varied widely, and some figures initially identified as key players in the Union debate could simply not be located within the National Archives. Nonetheless, an attempt was made to balance the sample as evenly as possible, within the bounds of what could be located. Thus, works from the following authors were included; John Cockburn, George Lockhart, Henry Dundas, Andrew Fletcher, Sir Walter Scott and Alexander Rodger. This provides the corpus with one author from each side of the pro/anti-Union debate who was involved in setting up the Union (John Cockburn and George Lockhart), a politician from each side of the divide (Henry Dundas and Andrew Fletcher) and a creative author from opposing camps (Sir Walter Scott and Alexander Rodger), in order to identify creative use of Scots across the political spectrum.

Unfortunately, it proved to be easier to find the writings of anti-union authors than those who supported incorporation. This does not mean that there were no authors who saw the benefit and promises of the Union, and many continued to support its existence throughout the eighteenth century. Indeed some, such as James Buchanan, combined his support for the Union with orthoepist ideals concerning the Scots language. He argued that correct pronunciation and a shared language would strengthen the ties between Scotland and England significantly over and above the political union (Crowley, 1991; Dossena, 2005). However, tracking down the writings of such people, especially when limited to online searches and requests, has proven challenging. As a result, the volume of work produced by the political figures included here is somewhat skewed towards the reactionary side of the political spectrum, although efforts have been made to reduce the effects of this where possible. For a full list of works see Appendix 1.

4.1.1.2.1 The Political Authors

John Cockburn (-1758) was a member of the Scottish and British parliaments and was actively involved in setting up and passing the Union agreement. Cockburn was strongly anti-Jacobite, and despite occasionally voicing concerns over the validity of the Union, he remained overall a staunch

supporter. After the Hanoverian succession he actively suppressed attempts to dissolve the Union (Wilkinson, 2002). Cockburn can thus be considered a pro-Union supporter, placing him on the loyalist side of the political spectrum.

George Lockhart (1673–1731) was one of the commissioners in charge of organising the Union.

Despite this, he was strongly against the Union and the bribery involved in its transactions, working as an informant for the Jacobites with whom he had sympathies (Scott, 1992; Szechi, 1997).

Although Lockhart firmly opposed a constitutional union between England and Scotland, he was open to the idea of a closer, federal union (Scott, 1992). He was not inherently anti-institutional, but rather disagreed with the absence of Scottish representation in the Union transactions by landed families. (Szechi, 1997). Lockhart took part in an unsuccessful attempt to repeal the Union and was deeply implicated in the Jacobite Rising of 1715 (Szechi, 1997). Lockhart also forms an interesting linguistic case. Although his political sympathies were clearly anti-Union, he was also in charge of organising the political agreement, which would have required a particular social and linguistic conduct.

Henry Dundas (1742–1811), First Viscount Melville, was a Scottish advocate and Tory member of the Scottish Parliament. Dundas became extremely skilled in managing the Scottish parliament, his time in office saw a number of major accomplishments, including the abolition of slavery, the domination of the East India Company and the prosecution of the war against France (Fry, 1992). He was also a powerful and dominating figure, obtaining almost complete control over the Scottish parliament which earned him various nicknames, including “The Great Manager of Scotland”, “The Great Tyrant”, “King Harry the Ninth” and “The Uncrowned King of Scotland” (Matheson, 1933). He was virulently anti-Radical and was in a constant battle to abolish the radical movement. He also developed a long-standing and trusting relationship with the English Prime Minister, William Pitt the Younger (Furber, 1931). Yet despite Dundas’ anti-radical efforts, clear pro-Union stance and favourable relationship with English politics and the Prime Minister, he was also a proud Scots

speaker and distinguished himself with his argumentative and colourful speeches (Fry, 1992). It will be interesting to observe whether this regional pride in speech will carry over to his writings, or whether the political leanings of Dundas will encourage his language choices to follow along pro-Union lines. Nonetheless, his career and background clearly mark him as a pro-Union politician for this time period.

Andrew Fletcher (1655-1716) was a notable opponent of the Union and became widely recognised as an independent patriot and prominent opposition speaker (Cannon, 2015). Fletcher had a deep mistrust of the royal government and resolutely opposed any arbitrary actions on the part of the English Church of State in Scotland. His concerns included limiting the power of the monarchy, proposing an independent parliament and establishing frequent elections in order to limit the clear bribery that took place (McClean & McMillan, 2009; Scott, 1992). He sought to protect Scottish nationhood by arguing against the proposed 'incorporating union', pushing instead for a federal union (Scott, 1992). Although ultimately unsuccessful, one of his most famous contributions to the debate were his "twelve limitations", intended to limit English power in Scottish politics. These resolutions did not pass, but the *Act of Security* that was eventually enacted was largely based on them (Scott, 1992). He wrote bitterly of the perils of incorporation and conquest, and the sacrifice it entailed for Scotland;

‘The Scots deserve no pity, if they voluntarily surrender their united and separate interests to the mercy of a united Parliament ... in this trap of their own making’

State of the Controversy betwixt United and Separate Parliaments (1706)

After the Union took place, Fletcher, disappointed with the outcome, left politics and Scotland to pursue other interests elsewhere (Scott, 1992). His anti-Union sentiments thus place him securely on the opposition side of the political spectrum.

Sir Walter Scott (1771-1832) was a Scottish novelist, poet, historian, and biographer with a deep interest in the historic struggles characterising Scotland's past. Scott is characteristic for his dual character and beliefs – he was both captivated by the glamour of Scotland's violent and heroic past and simultaneously a firm believer in reason, moderation and commercial progress (Daiches, 1971). Just as there were opposing forces at work in eighteenth century Scottish society, so Scott reflected diametric interests concerning the international, refined and progressive, and the local, popular and traditional. Indeed, it seems 'his head belonged to one, his heart to the other' (Daiches, 1971: 43). Scott grew up listening to the tales and songs of the Jacobite Rising and one of Scott's best-known novels, *Waverley* (1814) was a reinterpretation of the Rising of 1745 and the lost way of life that had once characterised the Scottish Highlands (Daiches, 1971).

Yet Scott was very much a man of the Enlightenment. He championed tolerance and moderation, deplored the French Revolution and its aftermath and believed soundly in hierarchy and the peace that a stable power structure could bring, despite its costs (Wagenknecht, 1991). Scott's mixed reaction to the Union of 1707 is therefore unsurprising. Although he welcomed the Union, seeing it as a promise of economic prosperity and modernisation for Scotland, he also bitterly mourned the loss of independence, and felt Scotland's sense of national identity and tradition to be dying (Daiches, 1971).

Many of Scott's poems and novels combine his vast knowledge of Scottish history and society, his antiquarian interests and his romantic interpretations of Scotland's past, with his understanding that Scotland's interests were inextricably tied to a British future (Daiches, 1971; Wagenknecht, 1991). Scott saw both the strengths and weaknesses the Union, and was at once nostalgic and romantic, but also pragmatic and progressive (Wagenknecht, 1991). Scott was also talented in dialect shifting, able to express himself equally with eloquence and force in Scots and in polished English. Thus, Scott is not clearly situated on either side of the divide. The moderate and balanced outlook that

characterised Sir Walter Scott's political persona, and his skill in moving across the linguistic continuum could lead to an interesting output in his writings.

Alexander Rodger (1784-1846) was a poet and songwriter, becoming popular in radical literature and writing frequently in satirical broadsides (single-sheet newspapers). He also published satirical material and radical pieces in sympathetic newspapers such as *The Glasgow Reformer* and *The Spirit of the Union*, of which he was later editor and subeditor respectively [National Library of Scotland]. Rodger was anti-royalist and identified with radical sentiments, utilising the medium of poetry and song to make pointed political comments. He directly parodied the loyalist *Carle, now the King's Come*; a work by Sir Walter Scott produced especially for the royal visit of George IV to Edinburgh, with his own, *Sawney, now the King's Come* [National Library of Scotland]. Rodger's open dislike of English rule and domination therefore places him well on the opposition side of political affiliations.

4.1.1.2.2 The Political Documents

The documents chosen to be included from these authors include the online, digitalised Sir Walter Scott Correspondence spanning the years 1787-1832 (<http://www.walterscott.lib.ed.ac.uk>), two books by George Lockhart: *Memoirs concerning Scotland, 1707-1708*, and *Memoirs concerning the affairs of Scotland*, both which have been digitalised and are readily available online (https://archive.org/details/bub_gb_XA8-AQAAMAAJ), Andrew Fletcher's: *An historical account of the ancient rights and power of the Parliament of Scotland. To which is prefixed, a short introduction upon government in general*⁵, which is also available online as a digitalised book (<https://archive.org/details/anhistoricalacc00ridpgooq>), and select pages from works by John Cockburn, Alexander Rodger and Andrew Fletcher (for a full list of works see Appendix 1).

⁵ Although it is generally accepted that Andrew Fletcher wrote this book, it did undergo editing under George Ridpath (d. 1726) and was only published in 1823.

Unfortunately, sourcing a large quantity of digitised texts and manuscripts produced by these authors was simply not possible. As mentioned before, there is currently no repository of politically-based, historical Scots texts, and the correspondence of historical figures are often held in special or private collections, or are not available to the public. Through targeted searches for texts written by these six authors across the Scottish National Library, the National Records of Scotland and the Scottish National Archives, I was able to locate a number of texts across different holdings. However, some could not be accessed, or were for physical viewing only. Furthermore, some portfolios consisted entirely of handwritten material which is problematic for digitisation purposes (discussed further in section 4.1.2). Although handwritten texts are potentially a rich source of linguistic data, the time-consuming exercise involved in digitising them meant that such written documents had to be kept to a minimum in this study.

Some texts that were originally located in archive holdings were able to be sourced electronically elsewhere, such as the books by George Lockhart and the treatise by Andrew Fletcher. However, the extent to which these electronic versions are true to the original copy is not always clear. This is an issue with edited versions both contemporary and current; the editing practices of the publisher are often unknown, and it remains guesswork as to how well these preserve the stylistic characteristics of the author. Both in the case of works published during the eighteenth century, and in online editions available today, the audience is almost always English and so the chances of anglicisation are high. Yet digitised versions are not necessarily any more anglicised than their edited, eighteenth-century counterparts, and for the purposes of a corpus study, a digitised version of a historical text can save a significant amount of manual work. Accordingly, the digitised texts mentioned above were briefly analysed to explore their linguistic content. They indicated that at least some Scots words were included in the online version, suggesting that the editing practices were at least partially true to the original. The correspondence of Dundas was sourced from the National Records of Scotland, while the remainder of the texts used for this study were sourced entirely from the National Library of Scotland (NLS).

Although the NLS contained works by many of the above authors, these works consisted of portfolios numbering in the hundreds of pages. Ideally the entire portfolios would be scanned and included in this new corpus, in order to obtain as much linguistic data as possible. This in turn creates a comprehensive view of the linguistic choices characterising these individuals. However, time constraints and the costs involved mitigated this possibility. The high level of manual correction and analysis involved in digitising historical documents did not allow for extensive collections of texts to be processed. Furthermore, the prohibitive costs involved in having the manuscripts scanned and sent across the world (£1.20 (NZ\$2.30) per page), meant this was simply not feasible. The documents furthermore could not be viewed online, these were available for physical viewing and photographing only.

Instead, the staff at the NLS kindly agreed to photocopy five pages from the middle of each portfolio, ensuring that the pages were entirely covered in text. It was hoped that this would provide a reasonable if somewhat brief snapshot of the language used by these figures. This approach is of course not entirely unproblematic – five pages can hardly be taken to be fully representative of each person in question. There is a possibility that the pages chosen could happen to be different to the overall style of these authors, or discuss certain topics which inhibit or encourage use of Scotticisms. Nonetheless, this was all that could be undertaken given the temporal and geographical constraints on this project, and it was hoped the text files might still shed some interesting insights into the particular linguistic practices of political individuals. In the very least, this could allow for a low-level, small-scale quantificational comparison across documents, with the potential to expand on this in the future. With these limitations in mind, I requested and received five pages from each of the five authors; these were scanned and emailed as pdf files.

4.1.2 OCR

Unfortunately, once the document scans were obtained from the NLS, it became clear they could not simply be uploaded to the corpus. LaBB-CAT (Fromont & Hay, 2008), the corpus-building tool used for this study (more on this in section 4.1.3), can only process text files, yet the scanned pages were sent as pdfs. Accordingly, these had to be converted to text before they could be uploaded, using Optical Character Recognition (OCR) software. OCR essentially involves taking a scanned image of text and using software to distinguish pixel patterns within the image, which are then translated into alphanumeric characters (Blanke et al., 2017). Part of the nature of historical documents is their physical format; they exist as paper-based, analogous copies rather than digitised, machine-readable sources of text. In order to extract the linguistic information from the manuscript to the screen, some digital conversion is required.

In the case of handwritten texts this process becomes highly problematic, but even typed manuscripts can pose major problems for digitisation, due to their dated, fragile state that is so often the nature of historical documents. Historical texts are characterised by poor image quality, damaged or faded characters, thin or fragile paper, ligatures, historical spelling variants, unevenly printed characters (resulting from historical printing processes), fuzzy character boundaries where the ink has bled over time, paper degradation, discolouration, blotches, cracks, dirt, and bleed through from the following page (Bukhari et al., 2017). Yet, both commercial OCR systems (like Abbyy and OmniPage) and open-source programmes (like OCRopus and Tesseract) have traditionally been optimized for clean, contemporary texts rather than historical documents (Bukhari et al., 2017). The large-scale digitization of historical archives has different requirements to standard OCR engines, due to the complex layouts and untidy nature of historical documents. Yet in terms of the market for OCR, historical documents make up a relatively small proportion of the demand (Blanke et al., 2017).

Furthermore, a study undertaken by Blanke et al. (2017) which tested various commercial and open source OCR programmes upon a range of historical documents, found that none of the programmes could produce consistently good results across the various corpora used. Rather, some were better at dealing with certain types of texts than others, which further complicates the methodological choices facing those working with archive digitalisation (Blanke et al., 2017: 82). There have been advances made over the years in high-end OCR systems using Hidden Markov Models, Long-Short-Term-Memory networks (LSTM) and Neural Networks to train software to recognise and convert historical documents. These techniques have been somewhat successful in recognising text in both printed and handwritten form (Breuel et al., 2013; Doetsch et al., 2014; Simistira et al., 2015). However, the process of training such programmes to recognise historical texts is extremely time-consuming, and is complicated further by the huge variability of spelling practices in many historical languages.

Converting documents containing Scots is similarly problematic, as the majority of programmes do not cater for old, ancient, medieval and non-standard scripts (Bukhari et al., 2017). Standard English is the closest alternative to Scots, but a software programme attempting to fit English words to Scots spellings is liable to make errors. There have been developments towards creating a specialised OCR platform designed specifically for digitising historical texts, utilising recent advances. These include a number of small-scale pilot programmes such as anyOCR (Bukhari et al., 2017) and OWP (Blanke et al., 2017). Unfortunately, a lack of funding has resulted in their failure to be developed further and enter the general market (Mike Bryant, p.c.). Furthermore, even with increasing success rates such models can still produce large error rates and usually fail to convert historical correspondence (Fischer, 2012). This then requires manual checking and correction.

Thus, OCR is not always optimal for historical documents, but the alternative is to type out the entire manuscript by hand, which is exceptionally time consuming and also prone to human error.

Currently, most corpus builders will use OCR as a starting point, before manually editing and

finishing the digital conversion themselves. Accordingly, I took the same approach. As most of my documents were typed rather than handwritten, I attempted to use an OCR programme to convert these. Originally, these files were converted to text using the Optical Character Recognition function of Adobe Acrobat Pro. This had a reasonable success rate in recognising the text in some of the cleaner and better-preserved documents. Unfortunately, the rendering was not as clear for the texts suffering more damage.

It became apparent that Adobe would not be suitable to convert all the manuscripts, and a number of other freely-available, open-source OCR programmes were trialled including FreeOCR, SimpleOCR, WPS PDF to Word Converter, OnlineOCR and Microsoft OneNote. None of these were remotely successful. However, upon recommendation, the commercial audio-learning programme Kurzweil 3000 (<https://www.kurzweilededu.com/>) was trialled, and its OCR Scan and Extract function produced close-to-accurate renderings for the remainder of the texts. Traditionally used to aid non-visual learning, Kurzweil 3000 has a high-quality OCR package that can reproduce scanned documents with the exact layout and format as found in the original, to enable the text to be read aloud. The audio function of this software is obviously irrelevant to this study, but the strength and accuracy of the OCR machine was a significant advantage for the digitisation process. Alongside this, Kurzweil 3000 has a number of additional beneficial features, including multiple bilingual reference sources and translations to any Google supported language. This programme had greater success in recognising the printed text, including the Scots words present, and was particularly adept at converting the historical books and bound volumes.

However, both Adobe Acrobat Pro and Kurzweil 3000 refused to recognise the eleven samples of handwritten letters that were included in our study (see Appendix 2). Accordingly, these samples had to be typed out by hand. The OCR output of all the other documents was then checked and manually corrected. Some documents required a greater amount of editing than others, and careful attention was paid to the Scots words to make sure they were accurately converted. Once this had

been completed, these manuscripts, along with the electronic texts sourced elsewhere, were then uploaded to the corpus building tool used for this analysis; LaBB-CAT (Fromont & Hay, 2008).

4.1.3 LaBB-CAT

LaBB-CAT (Fromont & Hay, 2008) is a corpus compilation and analysis tool, accessible by browser and able to store text, audio or video files and other annotations. Although initially created in order to store searchable, time-aligned transcripts of video and/or audio recordings, LaBB-CAT lends itself equally well to textual corpora alone. It has various built-in tools and options available that can be easily applied to textual-corpus analysis and investigation. The 'layered' nature of the LaBB-CAT data structure comes with a number of predefined annotation layers, such as various word filters, linguistic representations and the CELEX database (Baayen et al., 1995 – discussed further in section 4.1.3.1). This enables the researcher to search across different layers of representation, including orthographic, phonetic and syntactic layers, within text or speech files, and filter the results accordingly (Fromont & Hay, 2008).

Thus, it is possible to undertake a search incorporating multiple linguistic layers simultaneously, allowing researchers to home in on the phenomenon in question. For example, it is possible to find in a corpus all verbs containing the morpheme <-ing> and the vowel [a] in Standard English. Searches can include the whole corpus or selected texts, and can extract lexical items, orthographic variants or particular syntactic structures. This layered structure can easily be extended to include other possible word-layers as appropriate to the study in question, which can be manually or automatically incorporated by researchers to store whatever extra information is desired. Each word within the corpus can thus have a number of representations in each word layer, which can in turn be used to filter results using the general search matrix.

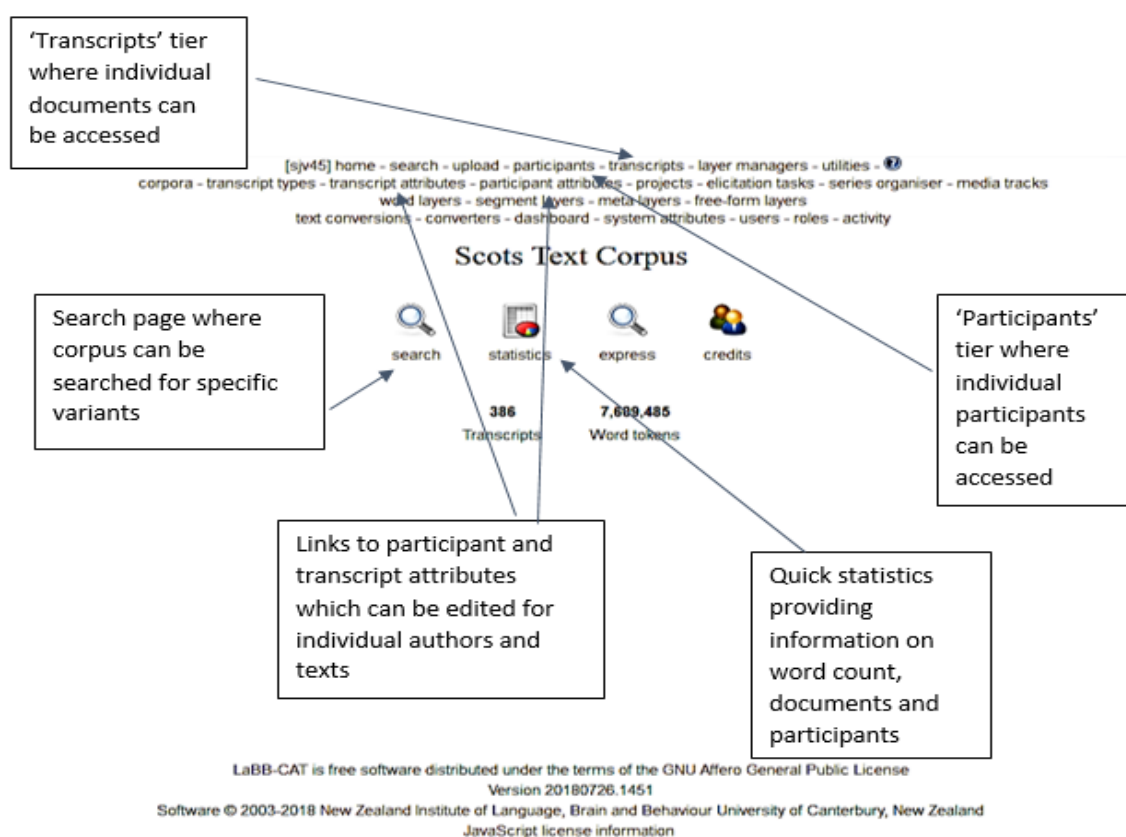


Figure 1: Home Page of LaBB-CAT for POLITECS

The search results or entire transcripts can be viewed or saved in a variety of formats, including a simple Comma Separated Values (CSV) format which can be directly exported to Microsoft Excel or other spreadsheet/database software programmes. Search results can be imported with various optional extralinguistic information about the speaker, the transcript, the full text of the sentence that matched the search pattern, and a URL for the sentence in the interactive transcript so that it can be accessed directly from the spreadsheet (Fromont & Hay, 2008).

Although there are various corpus-compilation tools available for linguistic research, the layered nature of LaBB-CAT gives it several advantages that were particularly beneficial for this study. These annotation layers enable the researcher to explore the corpus across various levels, in order to access specific information about the data contained within it. As well as the layers already included (such as orthographic, lexical, phonological layers (named 'pronounce' in LaBB-CAT)), the

ability to add or request additional layers, such as the CELEX annotation module mentioned above and non-standard frequency⁶ were able greatly assist a searching a corpus that contains two main languages; English and Scots.

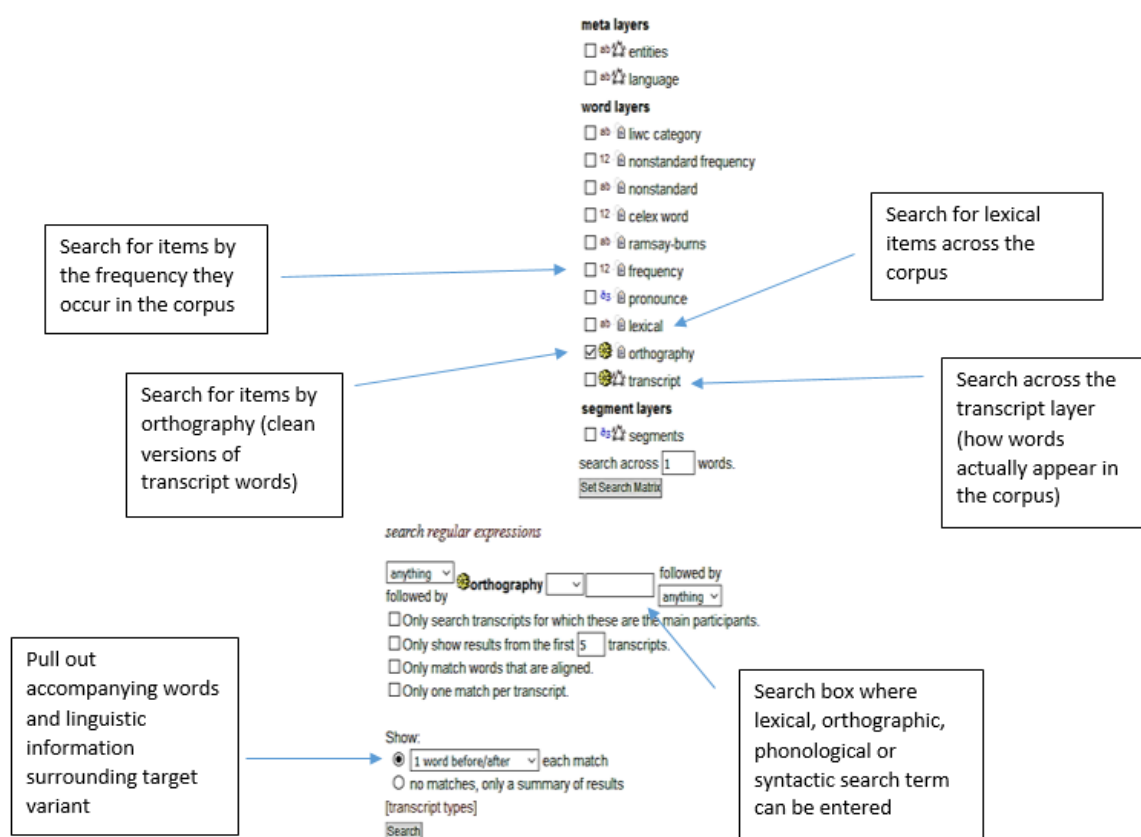


Figure 2: The search page of LaBB-CAT containing some of the various annotation layers for POLITECS

Furthermore, as LaBB-CAT has been developed at the University of Canterbury itself, this ensures a direct line of contact with its administrator and creator; Robert Fromont. This has enabled me to work closely with Robert to adjust and manipulate LaBB-CAT to suit my purposes, rather than being limited to the options and search levels built into a pre-existing corpus tool created elsewhere. The

⁶ The CELEX module provides access to the CELEX database (Baayen et al., 1995) which contains all Standard English lexemes, and the non-standard frequency manager identifies all non-standard English words. The properties and the use of these two filters is explained in greater detail in sections 4.1.3.1 (CELEX) and 4.2.1.2.1 (Cleaning the output).

changes initiated included incorporating a new word layer (in this case a Scottish word layer) managed by the LIWC (Linguistic Inquiry and Word Count) layer manager, a non-standard frequency filter, and the implementation of a text-only transcript converter for LaBB-CAT so that it ran more efficiently. More on these functions is discussed below. Robert was also able to quickly upload the digitally-available historical books used in this study through bypassing some of the validation steps usually needed for speech transcriptions, thus significantly speeding up the process.

4.1.3.1 CELEX

LaBB-CAT comes with CELEX incorporated into its structure as a normalised relational database, to facilitate looking up word information (Fromont & Hay, 2008: 8). CELEX is a database of Standard Dutch, German or English words (depending on which language is selected), along with the various linguistic classifications of each word, such as part of speech or its phonetic realisation. CELEX contains three distinct lexicons for each of the three languages; an abbreviation, a lemma and a wordform lexicon. The latter in effect contains all the words which are used in natural language (for example *walk*, *walking* and *walked* will be included as three separate entries in the English wordform lexicon (Burnage, 1990)) and it is this lexicon from the CELEX dataset that is incorporated into LaBB-CAT. Thus, every lexical item contained within the CELEX database is automatically available to researchers using LaBB-CAT.

When CELEX is included as a word layer during a search within the corpus, LaBB-CAT automatically generates an orthography layer along with the transcript layer. The words in this layer undergo a few transformations to tidy up the original text, to ensure the best possible chance of a match in the CELEX database. This largely involves removing all punctuation (except apostrophes and internal hyphens), enabling a better match success rate. Unfortunately, strange characters or symbols that may be present in the transcript can result in a mismatch. CELEX fails to recognise <j>, which commonly occurs in historical Scots documents (and indeed in various historical languages in general), so that the word <difuse> for example will not be recognised, although it represents the

Standard English word <disuse>. Non-alphanumeric symbols can also creep into text files as result of various text conversion and uploading processes that take place during the building of a corpus. For example, when historical documents are converted to text through the OCR process, a word such as <difuse> might be rendered as <di#!use> instead, which can slip through the net during the manual checking that follows the OCR process. Again, this word will not be recognised by CELEX.

However, Robert also included a non-standard frequency layer manager, which is able to access all the lexical items not recognised by CELEX. This includes tokens such as <difuse>, which can in turn be filtered and edited to reflect their true form once the results have been downloaded. By removing all standard punctuation characters from these non-standard words there is concurrently a much higher match rate with CELEX than by simply relying on the CELEX filter alone. The CELEX function was useful for the purposes of this study, by being able to identify which lexical items in our dataset were in fact Standard English words. The chances of Standard English words are considerably high given the long history of contact and borrowing in Scots (MacQueen, 1983) and the shared common lexical core between Scots and English (Meurman-Solin, 1993). By using the CELEX layer manager, Scots words could thus be separated from English words.

4.1.3.2 LIWC

Robert was also able to incorporate a Linguistic Inquiry and Word Count [LIWC] (Tausczik & Pennebaker, 2010) layer-manager into LaBB-CAT, enabling me to home in on the Scottishness of each text in the corpus. Traditionally LIWC is a text analysis model used to identify aspects of a writer's personality, by counting words in their writings that have been assigned to psychologically meaningful categories (referred to as 'dictionaries'). These include attentional focus, emotionality, social relationships, thinking styles, and individual differences (Tausczik & Pennebaker, 2010: 24). The LIWC programme is essentially based upon dictionaries, which refers in this case not to an alphabetical lexicon of a language, but rather a collection of words that define a particular category

(Tausczik & Pennebaker, 2010: 27). Multiple dictionaries can be created and run through the LIWC analysis one by one, to determine from which dictionary authors' select most of their words.

In my case I was not trying to tap into multiple aspects of the writers' personalities but rather a binary distinction between their 'Scottishness' or 'Englishness', and so I created just two dictionary files; a Scots dictionary and an English dictionary (more on the creation of these dictionaries is discussed in section 4.2.1). The LIWC programme is made up of two components. The first step processes the files that are fed into the LIWC layer manager, combing through the individual texts and comparing each word with the dictionary file it is provided with. Words are tagged when a match is found. The LIWC manager in LaBB-CAT thus tags words in the corpus with their category (Scots or English) according to the dictionaries provided. Each 'hit' indicates that a Scots or English lexical item contained within the dictionaries has been identified within the text. This is shown in Figure 3 below:

THE GORY PROFESSION. 1859. ATR 'Fy let us a' to the Bridal.'" A' ye that hae bumps o' destruction, Rejoice at the prospects o' war, Gae hire out yoursel as assassins, For Murder is yokin' his car. Fy haste ye to Mars and Minerva, And learn a' the throat-cutting trade, Get expert in the Gory Profession, And rob till your fortunes are made. CHORUS. Accoutre, and rush to the battle, Political murder's nae sin ; Its the Queen's highway to the devil ; Then, heroes, be loyal and rin. Should Russia's proud despot determine, To kindle the torches o' war, And grasp at our British dominions, He'll find baith a rock and a bar. Or if, for political reasons, Great Britain the Baltic should claim, Or seize upon Fez or Morocco, Then war, bloody war, is the game. Accoutre, and rush to the battle, &c. Come forward, political heroes, Enrol for the sea or the shore, Be ready for havock or carnage, The cause is the same as before, Just an honest crusade upon Freedom, To Queen and to country be true ; To kill and to murder's your duty, And so is the plundering too. Accoutre, and rush to the battle, &c. Gae sharp a' your tools for the battle, Rejoice at the cannon's loud roar : "Your glory, ye brave human butchers, Is wading knee-deep amang gore. How sweet to the true British sodger, Baith slaves and assassins by trade, Are the fields of brute-legal murders, Where havock and carnage are made. Accoutre, and rush to the battle, &c. What a noble profession is murder, Wheu sanctioned by King or by Queen ; Then might makes a rightful possession, Is truth that is legal I ween. Come forward ye sodgers and sailors, There's plunder in war's bloody game ; Ye

Figure 3: LIWC manager tagging a text with words identified from the Scots and English dictionaries

The second part of LIWC calculates the percentages of the variously-tagged words present in the texts. The module produces a list of the word-categories and the rates that each was used in a given text. Through counting the raw frequencies of Scots and English lexical items across all texts in the corpus, using the two dictionaries for reference, we can calculate the proportion of Scottishness in each text. The resulting output provides variable data that can be analysed using quantificational, sociolinguistic methods to determine which factors might be driving this variation.

Of course, this is a fairly rudimentary approach to truly estimating the ‘Scottishness’ or ‘Englishness’ of a text – there could be various syntactic constructions, semantic or pragmatic nuances and specific hybrid spellings that are indicative of Scots, but which we cannot access using lexical items alone. LIWC itself, like any computerized text analysis program, is problematic as a system due to the potential for miscoding or simply missing large chunks of valuable linguistic information in the signal. Such aspects of writing are difficult to explore using quantitative corpus methods as they often require a text-by-text analysis. This is part of the larger issue within corpus linguistics in general – the breadth and quantity of the texts we are dealing with simply mitigates the possibility for detailed individual studies into the idiosyncrasies of each author in question. Instead, by utilising current statistical methods, incorporating a large amount of data into the analysis, and analysing a select few texts in more detail, we can perhaps come a little closer to creating a more holistic understanding of language change in historical Scots.

4.1.3.3 Uploading


Once the OCR process was complete, the historical books, manuscripts and the text files from the CMSW were uploaded to LaBB-CAT. The text and participant information for the CMSW (provided as a master spreadsheet) was also incorporated into the new corpus. LaBB-CAT comes with a number of participant and transcript attributes built into its central database, and these contain the usual categories that tend to be analysed in sociolinguistic corpus studies; such as GENDER, SOCIO-ECONOMIC INDEX, AGE and TYPE OF SPEECH (*reading passage, interview, etc.*). However, with a bit of modifying

many of these attribute categories can be of use in a historical analysis. Categories such as GENDER, YEAR OF BIRTH, and social attributes are equally useful for historical linguistic research (Nevalainen, 1996, 1999, 2006; Nevalainen & Raumolin-Brunberg, 2000, 2003), whilst levels such as TYPE OF SPEECH and SPEAKER can be re-labelled as GENRE and AUTHOR respectively. Extra levels can also be added into LaBB-CAT for both Author and Manuscript attributes, so that categories such as PLACE OF PUBLICATION, YEAR WRITTEN, AUTHOR'S TITLE and POLITICAL STANCE could be included as well.

Accordingly, once the labels had been renamed and added to the corpus attribute structure, the participant and manuscript data was incorporated. LaBB-CAT is able to match the document ID of the texts to that supplied in the supporting documentation (the master spreadsheet), enabling quick and easy transferal of meta-linguistic data to the appropriate texts. Throughout the uploading process, LaBB-CAT consults all participant names to see if previously uploaded transcripts feature the same author, thus if a participant has produced multiple transcripts these are automatically grouped under the same author (Fromont & Hay, 2008). Once the textual transcripts have been uploaded to LaBB-CAT, additional information about the authors and text itself can also be stored. The attributes of a speaker can be accessed from the 'participant attributes' tab on LaBB-CAT's home page and selecting a particular author's attribute file. This is shown below in Figure 4; here we can see the attributes of Sir Walter Scott:

participant

Name:	Scott, Sir Walter
Gender:	Male
Birth Year:	1771
Notes:	General notes
Author's title:	Sir
Forenames:	Walter
Surname:	Scott
Initials:	Initials
AKA:	AKA
Confidence in year:	5
Place of birth:	Edinburgh, Scotland
Mother's place of birth:	Mother's place of birth
Father's place of birth:	Father's place of birth
Occupation:	Author, solicitor
Education:	University
Mother's Occupation:	Mother's Occupation
Father's Occupation:	Solicitor
Locations where resident:	Edinburgh
Other languages spoken:	Latin
Religious affiliation:	Religious affiliation
authorID:	47

 Save

participant's corpora
[pass phrase]
all utterances






cm5w-0189-y3-g3-Letter-from-Scott-to-Gifford-25-Oct-180.txt  Delete  Main Participants  Attributes  Generate  Media

Figure 4: Participant attributes for example participant (Sir Walter Scott) in LaBB-CAT

The e-books and remainder of the texts sourced from elsewhere required their meta-data to be added in manually, and this information can be included during the uploading stage in the appropriate transcript and participant attribute tiers.

The corpus-building stage was thus complete. A pre-existing corpus had been identified and uploaded to LaBB-CAT to fulfil the general component of the corpus, along with the corresponding extra-linguistic information. For the political component a number of politically-active authors and their texts had been identified, located, converted and uploaded to this new, custom-built corpus. This enables the comparison between how politically-motivated authors were using Scots in relation to their non-political peers. LaBB-CAT was then modified slightly to activate various filters, such as the CELEX database and the LIWC manager, to assist in corpus searches and develop a strategic system to tag Scotticisms in texts. The next step was to circumscribe the variable and then search the corpus itself.

4.2 The variable phenomenon

Although quantitative research into Old and Middle Scots has increased since the earliest accounts (Bald, 1926, 1927, 1928; MacQueen, 1957), analyses of eighteenth-century Scots have remained largely descriptive in comparison (Millar, 2004; Murison, 1979; Aitken, 1984; Jones, 1995; Robinson, 1973; McClure, 1980; Beal, 1997; Dossena, 1997; Smith, 1996, 2007). Research on Middle Scots has progressed from descriptive statistics (Devitt, 1989a; Meurman-Solin, 1989a, 1989b, 1989c, 1992, 1993a, 1997a, 2000b, 2003a), to statistical analyses incorporating a sociolinguistic methodology (Romaine, 1982), although there is still a need for the modern statistical methods frequently used in the analysis of contemporary corpus data (Hay et al. 2015; Gries 2016)⁷.

In order to undertake such an approach, a corpus is required, and a greater number of features need to be examined simultaneously. Accordingly, I attempted to undertake a more holistic approach to understanding language change in eighteenth century Scots. Unlike previous studies which have focussed largely on examining single orthographic features or the raw frequency of Scots lexical items (see Cruickshank, 2012; Corbett, 2013), this study sought to statistically analyse a large number of Scots variants simultaneously within a corpus of texts.

This is where the strength of the LIWC analysis comes into play. This allows the researcher to examine hundreds of features simultaneously, by compiling an untold number of lexical items under a single category (in this case the category ‘Scots’ or the category ‘English’). Rather than providing a raw frequency count of all Scots words in the corpus, the Scots words can be condensed into a Scots dictionary, with a corresponding English dictionary. Creating two dictionaries which incorporate large numbers of lexical variants existing during the eighteenth and early nineteenth centuries allows for a simple binary distinction between ‘Scots’ and ‘English’, which is required for the next step of the process: the statistical modelling. Their frequencies can be quantified via a LIWC analysis

⁷ Although see van Eyndhoven & Clark (forthcoming) for a re-examination of the <quh-> variant in Middle Scots using current statistical modelling methods.

and then further explored using statistical tools to compare usage across texts and time. The first stage is thus to identify a large number of lexical items, and compile these into a dictionary. This process is discussed in section 4.2.1 below.

4.2.1 Dictionary compilation

In order to undertake a LIWC analysis, lexical items from both Scots and English were required, to assess how often authors chose options from one language or the other. These could not just be any Scots or English word; they had to be equivalents of one another. Particular Scots words and their English translation (where a lexical equivalent could be identified) were required, and once identified these could be added to their corresponding dictionaries. The English equivalent ensures we can identify not just when Scots words were used, but also when the anglicised choice was used instead. It is important to remember that Anglicisation had been going on for well over a century by this time, and many written mediums had incorporated a large number of English lexical items into their registers (Murison, 1979; Jones, 1993). Furthermore, Scots and English share a large common core of lexical items and spellings (Meurman-Solin, 1993).

Thus, to simply quantify the overall number of English words in each text would be missing the point. Some English lexical items present in the texts may have stopped being variable long before the eighteenth century. Clearly, there will be a higher proportion of English words than Scots words in the texts, but this is not to say that the instances where Scots was used are insignificant. Rather, we wish to determine how often a Scots lexical item or spelling was used instead of the English variant, *when there was variation*, to determine how often authors were *variable*. For example, in the eighteenth century the word *oak* was written in General Scots as *aik* or *ake*, thus we would be interested in finding all instances of both *aik* and *oak* in the texts of the corpus.

Yet Scots words are similarly problematic, as not all of these necessarily have an equivalent. Some words can only be translated as a description rather than correlating to one particular word, such as

bergset which the DSL describes as ‘a rock on the sea-shore from which angling is carried on’ (<http://www.dsl.ac.uk/entry/snd/bergset>). Others may have multiple translations into English depending on the context of the sentence, for example, the DSL translates the Scots word *raff* (and derivatives *raffie*, *raffy*) as ‘1. plenty, abundance. 2. a large number, crowd. 3. thriving, healthy, flourishing. 4. rank growth, 5. coarse-textured and 6. worthless stuff, rubbish’ (http://www.dsl.ac.uk/entry/snd/raff_n1). It is clear these words do not have straightforward translations into English. Other lexical items may have no translation at all, being tied to a concept or aspect that is distinctly and inherently Scottish. For example, *Beltane* refers to the first or third day of May, and is one of the ancient quarter days of Scotland, during which a fire festival is observed on the hill-tops and occurs particularly, but not exclusively, in the Highlands (<http://www.dsl.ac.uk/entry/snd/beltane>). For such words, it is unlikely that authors would have varied, as they would be describing something that did not exist outside Scotland’s borders.

Accordingly, the Scots dictionary would have to contain only Scots words that were variable and had a single, straightforward English equivalent. To achieve this, I initially sought to simply scrape all entries from the online Dictionary of the Scottish Language (DSL) [<http://www.dsl.ac.uk/>], as this could effectively provide a ready-made wordlist that, once filtered and sorted, could be fed into the LIWC layer manager to form the ‘Scottish’ dictionary. However, when this was attempted it soon became obvious that the results could not be easily processed as a result of inconsistent HTML coding. The headwords and their translations or descriptions were marked in widely varying formats within the coding (such as parentheses, various alphanumeric characters or no marking whatsoever), making it impossible to filter the results by word or translation into some coherent form.

The option to use the DSL was complicated further by the problems identified with Scottish lexical items above (no equivalent existed, or too many definitions were given), as well as the high number of identical words across English and Scots, as a result of their shared parent language and the history of contact that characterises Scots (MacQueen, 1957). It is impossible to untangle from a

corpus-analysis alone whether the author was tapping into ‘Scottishness’ when using shared lexical items, or anglicising their work, and thus to include them in the dictionary can over-estimate levels of Scotticisation. Instead it is preferable to err on the side of caution and include only clearly Scots words in the analysis. This negates the possibility of utilising a pre-existing dictionary, such as the DSL, which contains both Scots and English words. Furthermore, the variable spelling options present within early Modern Scots are not always defined or categorised within the DSL. Whilst frequently appearing in examples or as abbreviated alternatives, not all hybrid forms of particular lexical items are individually listed, and thus to rely purely on dictionary entries would miss a substantial amount of the variation that exists within Scots. As a result, I sought to create my own, unique Scots dictionary, rather than relying on the DSL. By creating my own dictionary there was a lot more control over the items that were included, and importantly, did not include any lexical items that were undoubtedly English in origin.

4.2.1.1 Word Lists

Following on from this initial attempt, I sought instead to use pre-existing Scots wordlists to form the basis of the Scots dictionary. A large number of lexical items were taken from the wordlist provided by Corbett’s 2013 study, which analysed two poems by Allan Ramsay and Robert Burns. Corbett identified all the Scots words used in the poems, including their hybrid forms⁸. This list was quite extensive as a result of the creative and artistic nature of the texts it was sourced from, as well as being based upon the time period in question, making it a useful resource and starting point. Each word was located in the DSL to identify whether an English translation existed. As mentioned earlier, many Scots lexical items cannot necessarily be correlated with an English option. Thus, where there was a single, straightforward equivalent, both the Scots and English words were added to their

⁸ Corbett (2013) identified a number of emerging hybrid spellings present in the poems, and suggests that these arose during the eighteenth century in such creative genres to expand the variable language system, allowing them to reduce the unfamiliarity of Scots words to an English-speaking audience, or highlight a Scots pronunciation of a shared lexical item (p. 7). Often Scottish poets drew on Standard English orthographic practices when introducing Scots words, in order to draw difference from English and utilise the creative extension that Scots allowed them, whilst maintaining intelligibility (p. 65). Scots lexical items with these hybrid spellings are included in the wordlist.

corresponding dictionaries. If a clear English alternative did not exist, or reflected a unique aspect of Scottish life and thus can never be expected to vary, then the word was not included in the dictionary. Of course, this removes some potential sources of Scots, and it is possible that authors deliberately chose to use such items for a particular pragmatic or creative effect. However, without a full discourse-analysis of each text in the corpus, it is impossible to assess this possibility from token counts alone. Through removing these sites of non-variation, we are at least curtailing the focus only to truly variable lexical items, homing in on the variation that existed rather than codified markers of Scottishness. This process removed 52 words, leaving 282 items to be added to the dictionary.

The Scots dictionary was then expanded by drawing on other wordlists concerning Middle and Modern Scots lexical items, including those mentioned in Aitken (1979, 1984, 1990, and 1997), Agutter & Cowan (1981), Riach (1984) and Dossena (2005). In many cases an English translation was provided alongside the Scots wordlist, which enabled for the Scots lexical items to have an English counterpart. This boundary can be unclear occasionally, but the words included in the above research were often those that have attracted attention as a result of their salient nature. They frequently reflected the choice to use a Scots lexical item in light of the clear and relatively well-established English equivalent, suggesting that writers or speakers made some kind of choice, either consciously or unconsciously, between these two variants within their work or speech. This is ideal for the purposes of this study. Accordingly, applicable Scots lexical items were taken from these sources and added to the Scots dictionary, whilst their English translation, was added to the English dictionary. This added a further 227 words to each dictionary.

This provided a good starting point, but does not yet come close to the more holistic approach this research sought to achieve. Yet it is clear already that the binary analysis required by the LIWC remains problematic for lexical items even with rigorous checks in place. Choices are not necessarily binary and sometimes the English or Scots variant is simply inappropriate for the context.

Nonetheless, as this study seeks to utilise statistical modelling to uncover variation between Scots

and English, the envelope of variation had to be defined. I could not rely purely on Scots lexical items alone, and so instead orthographic variants – particular Scots spellings and their English equivalent - were chosen to form the variable basis of the dictionaries. Scots orthographic variants are in this regard somewhat easier to incorporate into a historical sociolinguistic analysis, as the distinction between the Scots spelling (or spellings) and the anglicised spelling is often less subjective than lexical items. However, a spelling variant cannot be uploaded to either dictionary on its own, as the LIWC dictionaries rely on lexical items. Instead, lexical items containing these spellings variants were included, along with particular words and word-lists. The process of identifying these is discussed below in sections 4.2.1.2 and 4.2.1.3.

4.2.1.2 Orthographic variants – Scots

Scots orthographic variants have the potential to encompass a higher proportion of words that were truly variable in early Modern Scots – these do not differ by their semantic meaning but rather their orthographic form. Authors writing during the eighteenth century could thus choose to spell the same word with the Scots or English orthographic variant, if the lexical item allowed for this variation. Utilising orthographic variants captures more of the variation present in the corpus, which in turn can be used to expand the Scots dictionary. The more comprehensive the ‘Scots’ dictionary, the more variation can be identified, allowing for a more robust analysis of the Scottishness of the texts in the corpus. The orthographic variants included in this study were taken from the modified Scots orthographic system identified in Corbett (2013). These are shown in the Table 1.

Table 1: Orthographic variants for early Modern Scots used for analysis, taken from Corbett (2013)

<i>Early Modern Scots</i>	<i>English</i>
<i>au+l</i>	<i>o+l</i>
<i>ane</i>	<i>one</i>

<i>ai, ay</i>	<i>oa</i>
<i>ei</i>	<i>ee</i>
<i>ee, ey</i>	<i>ie, ea</i>
<i>yi, ye, ie</i>	<i>y</i>
<i>ui, uy, u_e</i>	<i>oo</i>
<i>oo</i>	<i>ou</i>
<i>ch</i>	<i>gh</i>

However, accurately capturing all the lexical items that varied by these orthographic features in the corpus, required a two-step process. First, the corpus itself was searched for these particular orthographic variants. This provided an initial list of words containing the Scots spellings, and words containing the English spellings. These could then be checked and sorted, before being loaded into the corresponding Scots and English dictionaries, which were then loaded into the LIWC layer manager (this is discussed in more detail below). The second part of the process involved the LIWC manager taking these dictionaries and counting the frequencies of these words - taken from the corpus - in POLITECS itself (as well as counting the words added from the pre-existing wordlists, as discussed above). This second step provided the token counts, as well as the extralinguistic information attached to each token, which was required for the statistical modelling later on (this second step is discussed in more detail in section 4.2.2).

The first step was thus to search for each Scots spelling variant in the corpus, using the search string matrix built into LaBB-CAT. This generated a wordlist of all words in the corpus that contained the orthographic variant in question, which can be downloaded into a CSV file. However simply searching for certain orthographic variants across the corpus includes a lot of erroneous data and

does not accurately capture the Scots lexical items required. For example, although the spelling *auld* is identifiably Scots, searching for *au* in the corpus also generated Standard English words like *because* and *Paul*, and French words such as *aujourd'hui* and *beau*, and these have to be removed from the analysis. The resulting wordlists from these initial searches contained many English and French words among them, which made manually sorting them far too time consuming. However, the CELEX and the non-standard frequency layers in LaBB-CAT are jointly able to identify and filter out the Standard English words. The CELEX layer manager effectively tags all words generated in the search string as standard or non-standard by matching the lexical item with the target orthography contained within its database. It will only be able to match Standard English words and assign them a grammatical category. The non-standard frequency layer manager then selects all the lexical items that are not marked by the CELEX layer manager, and these non-standard words are presented in the results. Accordingly, Standard English words do not make it into the search results, whilst Scots lexical items remain unmarked and are thus included. These filters in LaBB-CAT are shown in Figure 5 below:

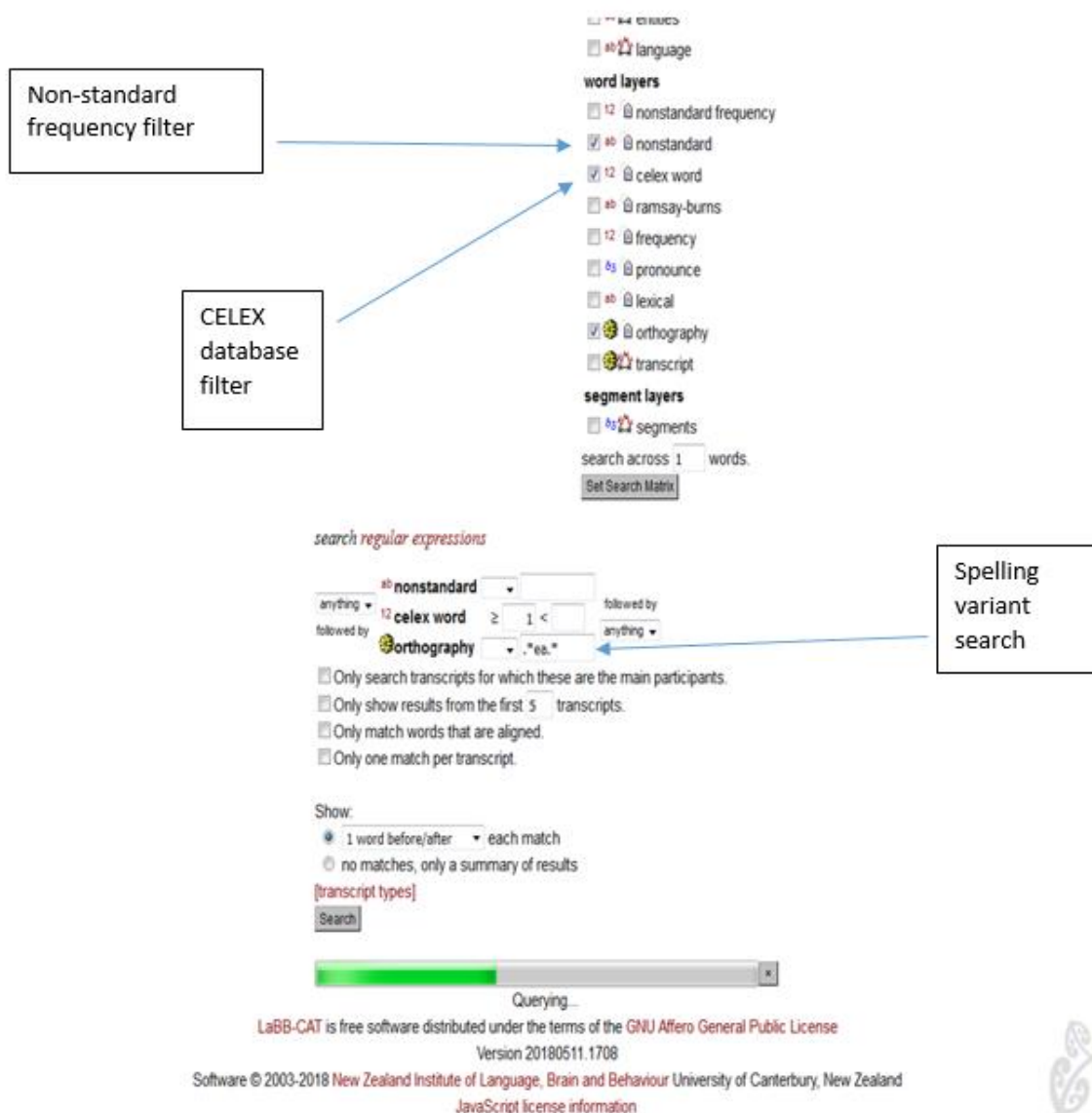


Figure 5: Various filters applied to a simple orthographic search string in LaBB-CAT which removes all Standard English words

Of course, this approach is not flawless; some Scots words have passed into Standard English over time, such as *skulduggery* and *wee*, and the layer manager may fail to recognise words that are historically English in origin, but which have become archaic or obsolete in Present Day English, such as *saule* (Middle English for *soul*) and *treillis* (various types of cloth) [<http://www.oed.com/>]. Nonetheless, this is the first step to circumscribing the results and sifting out the large numbers of

simply irrelevant data that we can safely classify as ‘not Scots’. Accordingly, these filters were checked before the Scots orthographic variants were searched in the corpus, and the resulting data then exported to a CSV file to undergo further cleaning.

4.2.1.2.1 Cleaning the output

Once the resulting datasets generated from each search were combined into a master spreadsheet, it became clear that the data still contained a large number of irrelevant and unclean tokens that needed to be checked and sorted. Although the lexical items obtained from each orthographic search contained the target orthography, and were generated from the corpus itself, this did not mean that all were Scots words. These tokens could not simply be added to the dictionary file, as this would cause the LIWC manager to tag irrelevant or inappropriate lexical items as Scots in the corpus. This would lead to over-reporting the number of Scots words in the corpus when the LIWC manager extracts the frequencies of Scots and English words from all texts.

The combined spreadsheet contained 43,991 tokens, with many incomplete words, page markings or titles, lexical items from other languages and strange renderings of words. A large chunk of this messy data occurred only once in the dataset, and thus I chose to remove all observations that only occurred once in the corpus. This brought the dataset down to 14,529 individual items. Although the single-frequency tokens altogether made up a substantial part of the dataset, it must be remembered that each individual lexical item only occurred once in the entire corpus, and a significant part of this consisted of the irrelevant tokens mentioned above. Removing them does eliminate some of the data we are dealing with, but their individual presence is relatively insignificant across the whole corpus. In the interests of time and efficiency removing such tokens spares considerable effort on the researcher’s part, especially concerning the processes involved in cleaning and sorting the data and running it against various databases (discussed further below). This approach is by no means ideal, but by this stage of the research there was little option available other than a quick and general clean of the dataset. The process of sourcing my own political texts,

converting them to text and building my own corpus, followed by various trials and errors in the attempt to source a larger number of variable Scots lexical items, meant that by this stage I was ten months into a twelve-month timeframe. Given more time, these tokens could certainly be cleaned more efficiently and checked manually, but this was simply not feasible for this research project. This can also spare significant time during the statistical analysis, as large data-sets can prove computationally intractable, even with modern hardware and optimal processing conditions (Tagliamonte & Baayen, 2012). Furthermore, this still left 14, 529 words to be added to the Scots dictionary, and these in turn will generate a much higher frequency of actual tokens, as there were multiple instances of these lexical items in the corpus.

There were also a number of Standard English words that the CELEX database had failed to recognise as a result of unusual characters or numbers immediately preceding or following the token, such as *letter1* or *#article*. CELEX also identified words with apostrophes as non-standard, such as *horse's*, and hyphenated words, such as *week-end*, despite most of these tokens clearly being English words. All unusual characters, including strange symbols, numbers and characters had to be removed from the dataset, to enable CELEX to then filter out the remaining Standard English words. To do this the dataset was read into the open-source, freely available, statistical programme R (R Core Team, 2013), and these features located with the `str_detect()` function from the **stringr** package (Wickham, 2015). These results were collated into a new data-set, from which the tokens containing apostrophes or hyphens were identified and loaded into a separate spreadsheet. These were manually checked, as to simply remove these characters would have generated unintelligible results, (changing for example *work'd* to *work d*) and in some instances the apostrophe distinguished a Scots from an English word, such as *pray't* (without the apostrophe this would become *pray*, and thus designated as English). Similarly changing *ee-broo* to *eebroo* would prevent the word from being located in the corpus, as its rendition in the dictionary would not match with the corpus. As there were only 85 hyphenated words, the English tokens in these were easily removed.

However, there were 1029 tokens with apostrophes. Manually checking them all would have proved time consuming, but including them in the Scots dictionary would overpredict the true percentages of Scotticisms within texts. When the results were sorted by apostrophe, a cursory glance over these tokens indicated that the vast majority of those with 's were English, whereas 'z seemed to be preferential as the plural form for many Scots words. Thus, I decided to delete all tokens with 's in them. This removed 521 tokens. Although this may have removed some Scots words, it seemed prudent to err on the side of caution and create a more conservative data-set than a liberal one. This still left 508 tokens with 'd, 't, 'r and 'z and these were checked manually using the DSL as a reference. The remaining tokens containing non-alphanumeric characters were uploaded to R (R Core Team, 2013) once more and all punctuation characters removed using the `str_replace_all()` function. These were fed back into the overall dataset.

Once this had been completed, the new dataframe was read into R (R Core Team, 2013) again, along with the second column of the CELEX database. Using the `match()` function in R, the dataset was compared with CELEX, which marked out the now-standard English words that were identifiable once the strange characters had been removed. The English words were subsequently deleted from the dataset. This still left many French words in the dataset. Accordingly, the same process was applied using the French equivalent of the CELEX database; *Lexique 3.8.2* (New et al., 2001). The first column of the Lexique database was loaded into R, and `match()` run again.

This identified all the French lexical items, however, many of these words were problematic as they are also considered Scots. For example, the Scots word *ait* can mean oat, or eat/ate, or a custom/bad habit (<http://www.dsl.ac.uk/results/ait>). *Ait* has a long history in Scots, arising in Old Scots and first appears in writing around the sixteenth century (as Scots for *oat* - <http://www.dsl.ac.uk/entry/dost/ate>). However, *ait* also exists in French as the third-person present subjunctive conjugation of the verb *avoir* (www.collinsdictionary.com/dictionary/french-english/ait). Due to the long period of contact with French as a result of the Auld Alliance, trade and religious

affiliations until the Reformation (see Murison, 1979), various French words have become part of the word-stock of historical Scots, and their use can reflect a Scots rather than French focus. Lexique identified 934 words that were potentially French, and these were manually checked using the DSL. French words that had passed into Scots before 1500, which is well before the time period under investigation, and which were or are commonplace, well-established Scots (as according to the DSL) were kept and labelled as ‘Scots’, the rest were subsequently deleted. Accordingly, once the apostrophes and hyphenated words had been manually checked and the English tokens removed, the non-alphanumeric characters removed and CELEX and Lexique 3.8.2 had been run again on the data, the remaining tokens that had not been filtered out through these various levels were fed back into the overall dataset. This left 11, 352 lexemes.

4.2.1.2.2 The Dictionary of the Scots Language

This left a clean, comprehensible dataset, however, not all the tokens were Scots. The dataset contained all ‘non-standard’ (i.e. non-CELEX) words, but this included Middle or Early Modern English words that are no longer current in Present Day English (as mentioned above) and unintelligible ‘noise’ (often the result of OCR processes involved in the compilation of the CMSW) that may stem from somewhere in the corpus itself. To simply upload this dataset would again grossly overpredict the levels of Scottishness within the corpus.

The next stage was thus to compare all these tokens with a list of Scots words and identify matching lexical items. To do so, the online Dictionary of the Scottish Language (DSL) was used. The URL of the DSL search page was fed into the `recursive()` function of R (R Core Team, 2013), and the information provided for each dictionary entry was scraped recursively by this function, and stored as separate text files. The resulting collection of files contained all the Scots words contained within the dictionary, their translation or meaning, and their example sentences. These files were bound into one large text file, and uploaded into R again along with the spreadsheet containing all the remaining non-standard items in the dataset pulled from the corpus. The spreadsheet and text file

were compared using `match()` – the words that were identified in both sets of data were kept. 1, 712 tokens were unable to be located within the information pulled from the DSL, and these were deleted. Most of these were irrelevant tokens, although the structure of the DSL can unfortunately prevent matches between the target lexical item and the entries contained within it. Sometimes alternative spellings are listed as whole words, which allows for a match, but in some cases only the variable morphemes of the target word are listed. For example, the word *speir* is listed with its alternative spellings *speer*, *spier* and *spear* whereas the word *Monanday*, is listed with its alternative spellings *-dy*, *Mona-*, *-in-*, *-on-*, *-un-*; *Munan-*, *-en-*, *-in-*, *-(n)on-*, *-un-*. The second entry would not therefore generate a hit for the word *Munanday* unless the alternative spelling is also listed in one of the examples, which is not always the case. Never-the-less, `match()` was fortunately able to locate 9, 640 tokens within the DSL. Although there is still some possibility for error, most of these lexical items can safely recognised as being used in the Scots language at one point or another, and do not represent English, French or any other language.

4.2.1.2.3 The Oxford English Dictionary

Finally, all words of Scottish origin used in the region of Scotland between the years 1100-1700, were downloaded from the Oxford English dictionary, along with their definition. The Oxford English Dictionary online enables researchers to filter results by region and language of origin, thus words that were Scottish in origin and use can be easily identified. The addition of these words was simply to increase the size of the Scottish dictionary file. The orthographic datasets created earlier, though containing a large number of Scots lexemes, were unfortunately reduced by the cleaning and filtering processes described above. Though this was a necessary process in order to eliminate the large amount of irrelevant data, some words that were Scottish in origin and passed into English over time, were removed in the process. Furthermore, although the words taken from the orthographic searches represent a large portion of the Scots lexical items available to authors, they are the modified orthographic variants that had undergone certain changes and anglicisation processes to generate their precise representation in eighteenth century Scotland.

However, this does not mean that authors could not occasionally turn to older spellings, or salient Scottish words that preserve aspects of the older orthographic system, in their writing. The use of an older Scots variant may be a deliberate choice on the behalf of the author, but such instances cannot be counted using the Early Modern Scots spelling system alone. Older words and spellings are of equal interest to the study, as they represent a section of the Scots lexicon that cannot be accessed easily via the orthographical searches, as well as representing a line of continuity between English and Scots. The word list provided by the OED was by no means exhaustive, as it contained only words that have passed into English over time. However, it was able to add to the dataset a few more lexical items of Scottish origin, that at one time or another did make it into the English language.

However, as it is problematic to simply label all the shared words as ‘Scots’, without undertaking a detailed discourse analysis for each text to determine whether the words really were being used in their Scottish sense or their English sense, some filters had to be applied again. The OED dataset was run against CELEX, which removed certain words and word combinations that are perfectly acceptable in Standard English, such as *High Church* and *Whitsunday*, but kept other combinations that are not registered in the CELEX database, and therefore probably words that entered certain English dialects, but not standard written English, such as *fastens-eve* (Shrove Tuesday).

To further filter the results, the OED words were run against an Early Modern English wordlist. This wordlist was created through a general search in the OED using the same technique as before, this time filtering results to all words used in literary English during the period 1700-1850. The Scottish OED words were loaded into R, along with the early Modern English dataset. Using the `match()` function in R (R Core Team, 2013), the Scots OED file was compared with the English OED file and all duplicates identified. These were then removed from the Scots dataset. The resulting dataset thus contained Scots words that were shared with English, but either entered the English language at a later date than the period under investigation here, or entered particular regional dialects or

colloquial registers of English, rather than standard literary English, which we would assume would be the target register for Scottish authors writing in English. Finally, the remaining words were examined along with their definition to identify whether a clear English equivalent existed. Where this was the case (as in the above example), both the Scots and English variant were included in the relevant dictionaries. If was not the case, the word was not included in the Scottish word_file.

4.2.1.3. Orthographic variants - English

Once all these processes had taken place, the Scottish 'dictionary' (to use LIWC terms) was now complete, and contained 16, 417 lexical items. This dictionary consisted of lexical items from the various word-lists mentioned above, the Scots lexemes identified from orthographical searches of the corpus itself (after the various filters had been applied) and the Scots words pulled from the Oxford English Dictionary.

The second stage of compilation was to create the corresponding English dictionary, to complement the Scottish one. A similar process was applied. The English translations from the word-lists (see Aitken, 1979, 1984, 1990, and 1997; Agutter & Cowan, 1981; Riach, 1984 and Dosenna, 2005) were added where appropriate, and search strings were generated in LaBB-CAT, this time for the English orthographic variants that corresponded to their Scots counterpart. Thus, for Scots <aul>, <ane>, <ai>, <ay>, <ei>, <ee>, <ey>, <yi>, <ye>, <ie>, <y>, <ui>, <uy>, <u_e>, <oi>, <oo> and <ch> the corresponding English equivalents , <one>, <oa>, <ee>, <ie>, <ea>, <y>, <oo>, <ou> and <gh> were searched within the corpus. However, this was not a straightforward process of compiling and downloading the results for the English orthographic searches. There were certain spellings that existed in both early Modern Scots and English but represented different vowel sounds. For example, <oo> occurs in both datasets, but in Scots it corresponds to English <ou> (as in *hoose* for *house*) whereas in the English dataset it frequently corresponds to Scots <u>, <ui> or <u_e>, (as in *gude/guid/gude* for *good*). This issue can be largely mitigated by downloading only the lexical items recognised by CELEX, but this ignores the Middle and early Modern English words that would have

been current during the eighteenth century but are no longer around today. Accordingly, results were obtained through a series of filters again. The first round of orthographic searches was undertaken using the CELEX filter, generating results containing purely Standard English words. This process is shown in Figure 6 below.

word layers

- ☐ 12 nonstandard frequency
- ☐ ab nonstandard
- ☒ 12 celex word
- ☐ ab ramsay-burns
- ☐ 12 frequency
- ☐ 43 pronounce
- ☐ ab lexical
- ☒ orthography
- ☐ transcript

segment layers

- ☒ 43 segments

search across 1 words.

search regular expressions

anything 1 followed by anything

☐ Only search transcripts for which these are the main participants.

☐ Only show results from the first 5 transcripts.

☐ Only match words that are aligned.

☐ Only one match per transcript.

Show:

☒ 1 word before/after ☐ each match

☐ no matches, only a summary of results

[transcript types]

Querying...

LaBB-CAT is free software distributed under the terms of the GNU Affero General Public License
 Version 20180511.1708
 Software © 2003-2018 New Zealand Institute of Language, Brain and Behaviour University of Canterbury, New Zealand
 JavaScript license information

Figure 6: LaBB-CAT's search page with various filters checked to enable a search for Standard English words only

4.2.1.3.1 Cleaning the output

The results generated for each variant were then run through `stringr()` to remove strange characters, and duplicates were removed in Excel. Finally, the remaining data was analysed manually to identify anomalies (for example words such as *Page2* and *Photocopied* are likely properties of the

corpus itself rather than the writings of individuals), as well as function words, which were deleted. To include words such as *do* and *as* would grossly bias the frequency counts for English tokens when the LIWC manager is tagging all the texts. Though these words may sometimes have varied, for most authors writing during this time period such words were no longer variable and thus not an accurate representation of the variation present in the corpus. Once anomalies and function words were deleted this left 8605 English lexemes.

Following this, the search-string in LaBB-CAT was modified, applying the non-standard frequency filter to each English orthographical search to find all non-standard lexical items for each spelling variant. This produced 79, 675 tokens and the results generated from these searches were similar in nature to the Scots spellings; there were large numbers of messy tokens, non-alphanumeric characters and lexical items from other languages, as well as many Scots tokens present within the dataset (especially for the shared orthographical variants). The same cleaning processes as the Scots dataset were applied using `stringr()`.

The remaining tokens were then compared with the CELEX and Lexique databases using `match()` in R (R Core Team, 2013). This time, the tokens generating a positive hit with the CELEX database were kept rather than discarded, which added 559 tokens. The French tokens were again manually checked against the OED, and 487 of these discarded, leaving 42 to be added to the English word file. The remaining tokens left over in the dataset were a mix of non-standard English words and Scots words, but the volume of results made manual analysis untenable. Instead, the remaining tokens in the dataset were loaded into R, along with the OED file of literary English words from the period 1700-1850, (created earlier when compiling the Scottish data). `match()` was run again and hits that came back positive were included in the final datafile, as these indicated a pre-Modern English word had been located. This generated 249 non-standard English words, which was combined with the remaining dataset. The resulting English dictionary file contained 14, 567 English lexemes and consisted of:

- Standard English words identified from the orthographic searches by CELEX
- Non-Standard English words identified from the orthographic searches by the OED
- The translation (where applicable) of the Scots words mentioned in various word lists (c.f. Aitken, 1979, 1984, 1990, and 1997; Agutter & Cowan, 1981; Riach, 1984 and Dosenna, 2005)
- The translation (where applicable) of the Scottish words located in the OED

4.2.1.4 *The Dictionaries*

The two dictionary files necessary for the LIWC analysis were thus complete. These dictionaries were then uploaded to LaBB-CAT as a new word-annotation layer, managed by the Linguistic Inquiry and Word-Count (LIWC) layer manager (Tausczik & Pennebaker, 2010). Thus, the first part of the two-step process was complete. A new corpus containing Scottish writing spanning the years 1700-1860 had been created, consisting of a pre-existing corpus (the CMSW) that provides information on the general trends of literate Scottish society, and of several self-sourced political texts that provides information on the trends of political individuals in eighteenth century Scotland. These texts could be searched for the Scottish or English words located in the two dictionary files and tagged accordingly, allowing me to quantify the levels of Scottishness in both general and political texts. This will be explored next. It is important to note that although the dictionaries created here stem from the corpus, they merely indicate which Scottish and English words are *present* (or might be present) in the corpus, but not the frequency with which they occur, nor any additional extralinguistic information (such as author, gender, genre, etc) accompanying each word. How the data was collected and categorised for the purposes of the statistical analysis, in order to answer the research questions, is explained in sections 4.2.2 and 4.2.3.

4.2.2 Extracting the variable

Once the dictionaries were uploaded to the LIWC manager, two search strings were generated in LaBB-CAT; one that extracted all words tagged ‘Scottish’ in the corpus, and the other extracting all tagged ‘English’ words. The results were extracted as CSV files, along with their accompanying extralinguistic information, including both participant attributes (such as gender, year of birth) and transcript attributes (genre, year of publication, etc).

LIWC layer manager checked

☒ @ liwc category
☐ @ nonstandard frequency
☐ @ nonstandard
☐ @ celex word
☐ @ ramsay-burns
☐ @ frequency
☐ @ pronounce
☐ @ lexical
☐ @ orthography
☐ @ transcript

segment layers
☐ @ segments

search across 1 words.
[Set Search Matrix](#)

search regular expressions

anything @ liwc category Scottish followed by anything

☐ Only search transcripts for which these are the main participants.
☐ Only show results from the first 5 transcripts.
☐ Only match words that are aligned.
☐ Only one match per transcript.

Show:
☒ 1 word before/after each match
☐ no matches, only a summary of results
 [transcript types]
[Search](#)

LIWC category set to Scottish to collect all 'Scottish' tagged words out of corpus

Accompanying sentence information (which can be set to 1, 5 or 10 words before and after target item)

LaBB-CAT is free software distributed under the terms of the GNU Affero General Public License
 Version 20180726.1451
 © 2003-2018 New Zealand Institute of Language, Brain and Behaviour University of Canterbury, New Zealand
[JavaScript license information](#)

Figure 7: Search string using LIWC category

Due to the high volume of textual information the process was computationally dense, and the initial results once obtained were sizeable. The ‘Scottish’ search indicated that words tagged as ‘Scottish’ occurred 209, 867 times in this corpus. The ‘English’ search, as can be expected, provided an even greater frequency of hits – 1, 590, 259 tokens were generated from the search. This is one

of the reasons a two-step process was required. Once the results were organised into spreadsheets it became clear that despite the rigorous cleaning applied to the dictionaries earlier, a few words had still slipped through the various filters. These were mostly surnames, and a number of high-frequency, function words, such as *of*. These lexemes were excluded from the dataset, leaving 867, 592 tokens.

4.2.3 Circumscribing the data

Following the extraction of Scots and English words, it became clear that a considerable amount of the extralinguistic information for the CMSW component of the corpus was missing, or coded as unknown. However, most of the missing information for these texts was able to be sourced through quick web searches, and this was accordingly incorporated. Certain factors were also recoded – for example there were instances of political prose in the CMSW that had been labelled as *Administrative Prose*. In POLITECS these were recoded to *Political – Prose*. The political leaning for each author in the corpus was added as well. This was coded as *Pro* for those supporting the Union, *Anti* for those against the Union, and *Unknown* when there was no information to be found on their possible viewpoint.

Once all extralinguistic information had been added in, the separate ‘Scottish’ and ‘English’ csv files were read into R (R Core Team, 2013) and bound into one overall dataframe using `rbind()`. This created a datafile with 858, 485 observations. As the CMSW contained texts dating from 1700-1950, there were a number of texts published in the last hundred years of this timeframe that were not relevant to the sociolinguistic component of this study⁹. Accordingly, texts published between 1707 (the year of the Union) and 1860 were kept in the dataset¹⁰. This left 519 texts to examine and 777, 423 tokens.

⁹ These texts were kept in the overall POLITECS corpus however, to allow for temporal analyses of the data to span more than just the eighteenth century, and to keep the possibility of future diachronic research across multiple centuries open.

¹⁰ This process required a careful text-by-text examination, as some of the documents in the corpus had in fact been written earlier on, but published at a much later date (in particular memoirs were published many years

4.2.4. Recoding factor levels

Once the data had been condensed to the appropriate timeframe, some further cleaning of the extralinguistic information was undertaken. Certain levels contained a large number of extralinguistic variables (such as BIRTHPLACE and PROFESSION) and this is problematic for any kind of statistical or quantificational analysis. It is difficult for any model to recognise patterns in the data and uncover meaningful relationships between predictors and the dependent variable if the data is thinly spread across multiple categories. Although tools such as random forests are able to work with highly imbalanced datasets more successfully, their predictive power is weakened by sparse sets of data and a large number of factor levels (Tagliamonte and Baayen, 2012).

Accordingly, several categories underwent a re-coding to aid in discovering trends in the data.

BIRTHPLACE was condensed to the three Scottish locations with the highest number of tokens (*Edinburgh, Glasgow, Aberdeen*) whilst the rest were combined under *Scotland_Other*, as these had very small proportions. Similarly, locations within England and Europe had very low counts, and thus were absorbed into the overarching categories *Europe* and *England*. PROFESSION also varied widely, and many of the participants in the corpus had multiple occupations during their lifetimes. As this project seeks to analyse the effects of political change on language, but also the language of those in politics, all participants who were politically involved at the time of writing were coded as *Politician*. For the non-political participants, their main profession was chosen, and these were grouped into four main occupations, with the remainder categorised as *Other*. This left *Politician, Author, Legal Professional, Poet, Orthoepist* and *Other*. MOTHER'S PLACE OF BIRTH and FATHER'S PLACE OF BIRTH had very little information, and so these were recoded simply as *Scotland, England, Other* and *Unknown*. Within GENDER, there were twenty-eight tokens listed as *Unknown*. As this was such a small percentage of the entire corpus, and such small counts can be problematic for statistical analysis,

after having been written). To aid in deciding which data could be removed, the author's Year of Birth was also checked. Authors born between 1650-1830 were kept – texts with authors born beyond this time frame were deleted from the dataset.

these speakers (two in total) were deleted from the results. The initial recoded values for each category are shown in Table 2 below.

Table 2: Table showing factors and the recoded factor levels for the POLTECS Corpus

<i>Predictor</i>	<i>Recoded as</i>
<i>Birthplace</i>	<i>Edinburgh</i> <i>Glasgow</i> <i>Aberdeen</i> <i>Scotland_other</i> <i>England</i> <i>France</i> <i>Unknown</i>
<i>Place Published</i>	<i>Edinburgh</i> <i>Glasgow</i> <i>Aberdeen</i> <i>Scotland</i> <i>England</i> <i>Europe</i> <i>Africa</i>

	<i>America</i> <i>Unknown</i>
<i>Profession</i>	<i>Author</i> <i>Poet</i> <i>Politician</i> <i>Legal Professional</i> <i>Orthoepist</i> <i>Other</i>
<i>Mother's Place of Birth</i>	<i>Scotland</i> <i>England</i> <i>Unknown</i>
<i>Father's Place of Birth</i>	<i>Scotland</i> <i>England</i> <i>Unknown</i>
<i>Other Languages Spoken</i>	<i>French</i> <i>Greek</i> <i>Latin</i> <i>Latin, French</i>

	<i>Latin, Greek</i> <i>Latin, Greek, Hebrew</i> <i>Multiple (3+ languages)¹¹</i> <i>None</i>
<i>Education</i>	<i>University</i> <i>Secondary School</i> <i>Boarding School</i> <i>Parish School</i> <i>Apprenticeship</i> <i>Unknown</i>

Some levels, such as LOCATIONS WHERE LIVED (all countries that the writer spent time in) and FATHER'S OCCUPATION were too problematic to easily standardise. Many authors moved around various locations during their lifetime, whilst FATHER'S OCCUPATION was often unknown. Although, with detailed research, it might perhaps be possible to pinpoint where authors were at the time of writing, or where they spent the greatest amount of time in their lives, or which occupation their father had while they were growing up, this is time consuming, laborious and the sporadic information available means clear trends remain unattainable. Thus, the FATHER'S OCCUPATION,

¹¹ The combination of Latin, Greek and Hebrew was common within the corpus, especially among academics, and thus this formed a category of its own. Multiple (3+) refers to authors who could speak three languages (though not the combination of Latin, Greek and Hebrew) or more. Those who spoke more languages could include Latin, Greek and Hebrew among them, but also an additional number of other languages.

MOTHER'S OCCUPATION (which was over 90% *Unknown*), LOCATIONS WHERE RESIDENT and RELIGION (78% of this was *Unknown*) were left out of the analysis altogether.

4.3 The data process revisited

Once this clean-up of the dataset had been undertaken, the final stage of the data compilation and collection was complete. A custom-built corpus had been created, using the pre-existing CMSW as a baseline and incorporating twenty-nine political documents sourced from the National Library of Scotland and online resources. The contents of the CMSW were directly uploaded into LaBB-CAT, whilst the archive manuscripts were converted to text using OCR before being uploaded. This formed a new corpus; POLITIECS. This corpus was searched using the various word-layer filters in LaBB-CAT to extract a list of all English and Scottish lexemes that contained target orthographic variants, identified as variable spellings in eighteenth century Scotland (c.f. Corbett, 2013). These were edited and sorted into two dictionary files - one Scottish and one English - before being uploaded to the LIWC layer manager in the corpus. Several pre-existing word lists, and a list of Scots words sourced from the OED were also added. The layer manager then tagged all instances of these words as they appeared in the corpus. Each occurrence of these words was extracted from the corpus, then sorted, cleaned and edited. Further extralinguistic information was added where necessary, before the tokens were combined into one overall dataset. Finally, certain predictor levels were recoded and modified to aid the statistical modelling.

The data was now ready to be explored temporarily and socio-historically. The particular trajectory of the Scots language during the eighteenth century could be examined, to determine firstly how the frequency of Scots lexis patterned over time for general literate Scottish society, and secondly how it patterned for politically-active individuals. Following this, I sought to determine which sociolinguistic factors were most important in conditioning the frequency of Scots lexis in general society, and which were most important among politically-active individuals, to assess possible correlations with the political and linguistic change occurring during the eighteenth century. Finally, with twenty-four

individuals in the corpus known to have been for or against the Union, I could explore whether their political sympathies potentially influenced their frequencies of Scots. These last three research questions require the use of current statistical methodologies and tools to explore the diachronic data. Before the results are presented, a brief overview of the benefits of statistical modelling is first given in the results and discussion chapter (section 5.1). This is followed by an explanation of the specific functioning of each of the different statistical tools used in this research, along with the corresponding results they generated, in the remainder of chapter 5.

5.0 Results and Discussion

5.1 The benefits of Statistical Modelling

The benefits of statistical modelling in linguistic analysis are by now widely attested, particularly in the field of quantitative sociolinguistics (Bickerton, 1973, 1979; Kay, 1978; Kay & McDaniel, 1979), and debate has since moved on to which method is best for modelling variable data. With recent developments and an increasing number of tools available in the statistical sociolinguistic toolkit, there are now ever-more nuanced and robust methods to analyse variable data, and potentially uncover new possibilities in data exploration, analysis and interpretation. Yet quantitative work within historical linguistics can be problematic, given the tendency for highly unbalanced data as a result of missing or empty data cells or inconsistent textual and author information; in other words, the ‘bad data’ problem (Labov, 1994: 11). Whilst modern variable analysis can work on balancing the number of tokens across different predictors to achieve a relatively stratified sample, historical analyses are left to work with whatever is available; usually resulting in incomplete and ambiguous data with a widely varying input across multiple predictor levels. Author information and their corresponding social traits, such as their sex or geographical location, cannot always be sourced from the texts themselves or located in historical records. Historical data can also span large collections, texts or time frames, exacerbating the problem of inconsistent information by distributing it thinly and unevenly across the data-frame.

However, with the tools now available the effects of this can be mitigated to some extent. There are a number of methodologies that can examine the effect of various extralinguistic variables upon a dependent variable within diachronic data, and these have already been applied to an increasing number of studies within the field of quantitative sociolinguistics. The tools of particular interest to this research include mixed-effects models, random forests, and conditional inference trees, which have not seen much application in Historical Scots (though see Smith (forthcoming)). Alongside extralinguistic constraints upon variation, the temporal patterning of the variation over time is also

of interest. Gries and Hilpert (2008) have developed Variability-based Neighbour Clustering (VNC) to interpret the chronological trajectory of a variant in a principled, data-driven way, and this technique will also be applied to this data.

Each of these methods provides a powerful and principled manner to examine the effects of social, linguistic and temporal factors upon the variation in the data, combining with and complementing one another to provide the researcher with a greater understanding of the dynamic and complex profile of a variable. Taken together, these data analysis methods can provide the most holistic look yet at any instance of language change, allowing the researcher to ‘go beyond the mere detection of a change and into the internal dynamics of that change’ (Hilpert & Gries, 2016: 3). Each method has its strengths and weaknesses; combining these tools provide new opportunities to greatly enhance our understanding of the historical process (Gries & Hilpert, 2010).

Accordingly, for this study I sought to utilise a combination of statistical modelling tools and methodologies to explore the variation in the data both in a temporal sense – to answer research questions one and two - and in relation to various socio-historical factors – to answer research questions three, four and five. A detailed explanation of how these tools operate, and how they were applied to my data, is explained further in section 5.2 below. The rest of the results section is structured as follows; first I shall examine how the frequency of Scots lexis patterned over time for the general literate Scottish society, and for politically-active individuals, using Variability-Based Neighbour Clusters (section 5.2). Then I shall explain how conditional inference trees and random forests function (section 5.3.1 and 5.3.2), before utilising these methods to examine which sociolinguistic factors were most important in influencing the frequency of Scots lexis in general society (section 5.3.3). This will be followed by an examination of the sociolinguistic factors influencing the politically-active members of POLITECS, using the same techniques (section 5.3.4) and finally, an exploration into the effect of political affiliation upon the linguistic choices of these individuals (section 5.3.4.2).

5.2 Statistical Modelling – The Temporal Analysis

5.2.1 Variability-Based Neighbour Clustering (VNC)

The first stage of the analysis was to use Variability-based Neighbour Clustering [VNC] (Gries & Hilpert, 2008) to examine how the frequency of Scots lexis patterned over time in both the general and politically-active Scots society. Gries and Hilpert (2008) have developed this hierarchical clustering model as new, data-driven approach to understanding diachronic data and how it patterns across time. VNC focusses purely on the temporal arrangement of variants rather than social or linguistic factors influencing them; it allows the researcher to explore the frequencies of a binary variable by indicating the concentrations of a variant along a linear time-line. Unlike other hierarchical clustering models, VNC recognises that the data is temporally constrained along a chronological timeframe. The model searches for clusters of similar data but simultaneously takes into account the temporal window of each cluster. Traditionally, diachronic data tends to be sectioned into convenient year-frames, yet this can disguise, ignore or alter trends, turning points and slopes in the data (Gries & Hilpert, 2010), as well as masking non-linear developments which may in turn discourage research across these convenient boundaries (Nevalainen, 2006). Temporal periodization of historical data is also often based on well-established time frames that have been defined by key socio-historical changes (Gries & Hilpert, 2010). Yet these time periods will not necessarily fit every variant being examined, and may miss crucial linguistic developments or ignore the time lag that may ripple through language change (Gries & Hilpert, 2010).

Furthermore, in this case there may be a potentially undiscovered factor influencing linguistic change (political change and unrest during the eighteenth century), as this study seeks to discover. If this factor is indeed discovered to be important, it is plausible that the second half the eighteenth and the beginnings of the nineteenth centuries in particular may behave differently to the decades on either side. Yet diachronic analyses which examine change in neat, hundred-year partitions could miss this effect. Pre-set year frames do not have the flexibility to allow for newly discovered factors

or influences, that may have operated broadly or specifically within the temporal space of the change under investigation, to be incorporated into their periodisation. Moreover, sectioning the data according to major historical events ignores the time lag that may ripple through language change, as well as ongoing underlying changes that may be present (Gries & Hilpert, 2010). VNC thus provides an alternative to artificially imposing pre-defined, fixed-length year frames upon the change in question, by grouping the data according to how it clusters over time.

The VNC script (which was kindly sent by Stephan Gries (p.c)) provides a principled method to determine coherent temporal stages as well as conservatively identifying data points as outliers. Firstly, the clustering algorithm takes as its input a data frame containing the frequency of Scots and English words per year between 1700 and 1900 the frequency of each variant. This then determines where along the timeline the variant (Scots) clusters most closely together. Clusters are defined by a high level of within-group similarity (standard deviation between the data points is lowest) and low level of across-cluster similarity (standard deviation between this group of data and another is suitably high), and this measurement for similarity can be altered so as to identify clusters that constitute a relatively homogenous period of interest. The script suggests an optimal range for the number of clusters to be included in the analysis, which is shown on a scatterplot. In the second stage the script produces a dendrogram, by overlaying the identified clusters on the data points, which are arranged along the timeline, as well as plotting the standard deviations between each datapoint. The result is temporal divisions of the data that are derived directly from the phenomenon under investigation (Gries & Hilpert, 2010).

5.2.2 VNC - The general literate Scottish population

Accordingly, I sought to use VNC to explore the overall frequencies of Scots words (relative to English) during the eighteenth century. This might suggest particular trends or turning points that could possibly stem from various historical factors. To obtain a holistic, overall sense of how Scots

patterned across general literate Scottish society, the frequency counts for Scots and English words were extracted for each year within the timeframe 1700-1893. This timeframe is larger than that under investigation, and this was done in order to compare Scots usage during the eighteenth century with the following century. This might enable us to see any marked differences between the two, as well as allowing for any potential time lag following events occurring around the turn of the nineteenth century. To further improve the results and aid with visualisation, years with less than five observations were deleted. This provides a clearer picture of what was happening in the language over time. The resulting data was then run through the VNC script in R (R Core Team, 2013). The model identified thirteen main clusters in the data¹², and these were accordingly plotted on the dendrogram shown in Figure 8 below.

¹² The model did identify a number of other clusters in the data, but, when plotted on a scatter plot, it became clear that most of these consisted of a single datapoint. These data points were widely spaced rather than tightly concentrated in defined temporal locations, which is indicative of outliers or extreme values within the data rather than concentrations of multiple points. This is suggestive of highly fluctuating data with inconsistent variability across time, rather than clear, defined trends, which is what we could expect given the competing forces at play during the eighteenth century. A number of other clusters were trialled, but the results were very similar. If more clusters were specified these were applied exclusively to the outliers, whilst a lower number of clusters simply condensed the main body of data, and similarly focus upon extreme values.

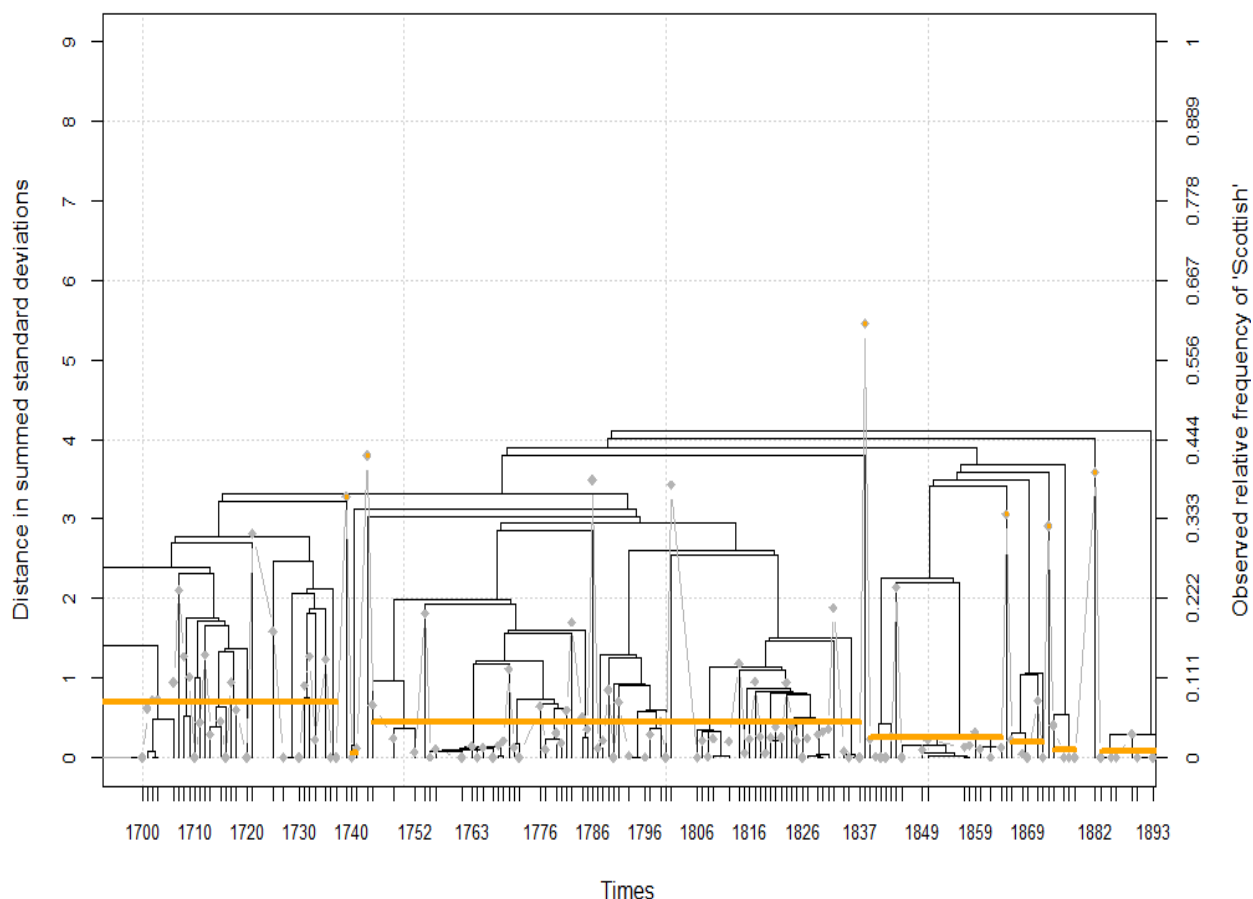


Figure 8: VNC analysis showing entire corpus with thirteen clusters

The yellow lines and dots indicate the clustering of the data, each line represents a number of years that behave similarly with respect to one another, and are dissimilar to the next data point along the time line. The x axis shows the timeframe for this data (the years 1700 – 1893). The left y axis indicates the difference in standard deviation between each of the merged temporal files; this reflects the level of deviation between the years contained within the clusters identified by the script. The higher the deviation, the greater the difference between one data point and the next. The right y axis shows the relative frequency of the Scots tokens. A higher value suggests a greater number of Scots tokens for that year relative to the overall number of tokens (Scots + English). These tokens can be observed as the raw data overlaying the graph – each point indicates the proportion of Scots tokens for that year. Points higher up indicate years with a higher level of Scottishness, points at zero suggest a very low proportion of Scots words relative to English words were recorded

for that year. The squares in the dendrogram show the mean standard deviations of each of the merged files – the larger squares indicate a greater level of deviation within the merged file.

As might be expected, there are not sudden, huge increases or decreases in the levels of Scots.

There are years with a sudden spike in Scots, but their presence suggests the effect of a few individuals rather than a dramatic, short-term increase in Scots. The overall proportion of Scottish words in the corpus is very low (about 8%) and thus the ratio of Scots to English words across each year is similarly low. When Scots levels are compared to overall frequencies (the sum of Scottish and English tokens for each year), the result is a very low frequency of Scots for this time period. The dendrogram does not simply plot the raw frequencies of all the Scots words – it does so relative to the English words. Thus, very high frequencies of English words combined with relatively low frequencies of Scots words will indicate almost no variation for that year. This trend is not surprising; Scots had been disappearing out of various written genres and professional arenas for well over a hundred years before this time period and we can see the continuation of this trend in the graph above. There is a small but steady decrease in the levels of Scots over time, and by 1870 word-usage is almost categorically English.

What is interesting to note is that the period 1744 – 1837 appears to be behaving somewhat similarly – this is one of the clusters identified by the model. These years are of particular interest as they match the time period when political and linguistic tensions were increasing as a result of conflicting interests and movements that both embraced and denigrated linguistic diversity and the established order. Though there was some turmoil immediately following the Union of 1707, aggravation only really began to build in the latter half of the eighteenth century as its effects begun to be felt. Similarly, the Augustinian culture that pervaded the early eighteenth century saw a rival movement in the shape of antiquarianism and the rise of vernacular Scots. The political dissatisfaction in particular carried on into the nineteenth century, to eventually fade with the rule of Queen Victoria (reigned 1837-1901) who did much to improve relations between the two nations

and inspired the romanticism of the highlands and Scottish history. While we do not see an abrupt increase in Scots, we do see a levelling off during this time period, and the number of large squares across the entire graph suggest that there was not a uniform decrease in written Scots across the board. Indeed, there is a greater difference in standard deviation in the second half of the graph, implying increasing variability over time. The result of this increased variation appears to be a plateau in the decline of Scots for the duration of this time period.

Considering the opposing forces at play, this is perhaps what we can expect. Although political and linguistic tensions were high, they were characterised by forces pulling in opposite directions – movements that embraced cultural assimilation and ‘polite’ English, and movements that championed the vernacular and independence. The result of this appears to have been a sort of equilibrium – Scots did not increase but nor did it continually decrease during this time. Although Scots usage clearly did slowly decrease overall, as is evident in the clustering shown here, the dendrogram does reflect a high level of variability throughout the eighteenth and early nineteenth century, and perhaps this variability can be in part attributed to the backlash to anglicisation and the Union. We can observe a noticeable fluctuation between individual years; the raw data on the graph oscillates between points of high usage (which suggest high levels of Scots words in texts produced during that year) and points that are at the x axis (which suggest categorical English usage in those texts). The considerable height and width of the squares (showing the mean standard deviations of each of the merged files) similarly indicates variability rather than uniformity – even within clusters there are considerable differences among data points.

In light of the various linguistic forces at work during this time period, such as the contradictory efforts of the orthoepists and antiquarians, as well as the various influences (including publishing pressures, codified legal and administrative terms and the linguistic scope available to creative writers) operating across the range of textual mediums and authors from various professions, perhaps this is not entirely unsurprising. Scots was declining in most professional arenas, with the

exception of legal and religious work which preserved archaic elements from the separate institutions of the Scottish law and church. Yet Scots also saw a resurgence in popular culture and creative works. These conflicting influences may explain why we see huge levels of variability from one year to the next. It is clear that the Scots language as a whole did not decline smoothly but jarred and jolted along the way.

5.2.3 VNC – the political members of the corpus

This provides an interesting first look at the frequencies of Scots usage across general Scottish society, and suggests that the eighteenth century - characterised by heightened political tension and an increasing linguistic awareness among the Scottish population - behaved differently to the decades on either side of it. In order to further unpick how political attitudes and stance affected levels of Scots, the data was subset to include only the politically-active people in the corpus. This included the six individuals that constitute the political component of POLITECS, as well as a number of authors identified within the CMSW as known political figures or espousing Unionist sentiments. This gave twenty-four authors in total for the political subset. Scots and English tokens from these individuals were fed into the VNC script, and a scree plot generated. This identified five clusters¹³, and produced the dendrogram shown in Figure 9 below.

¹³ Again, a number of outliers were identified in the clustering phase. As with the first VNC analysis (Figure 8), these were similarly ignored.

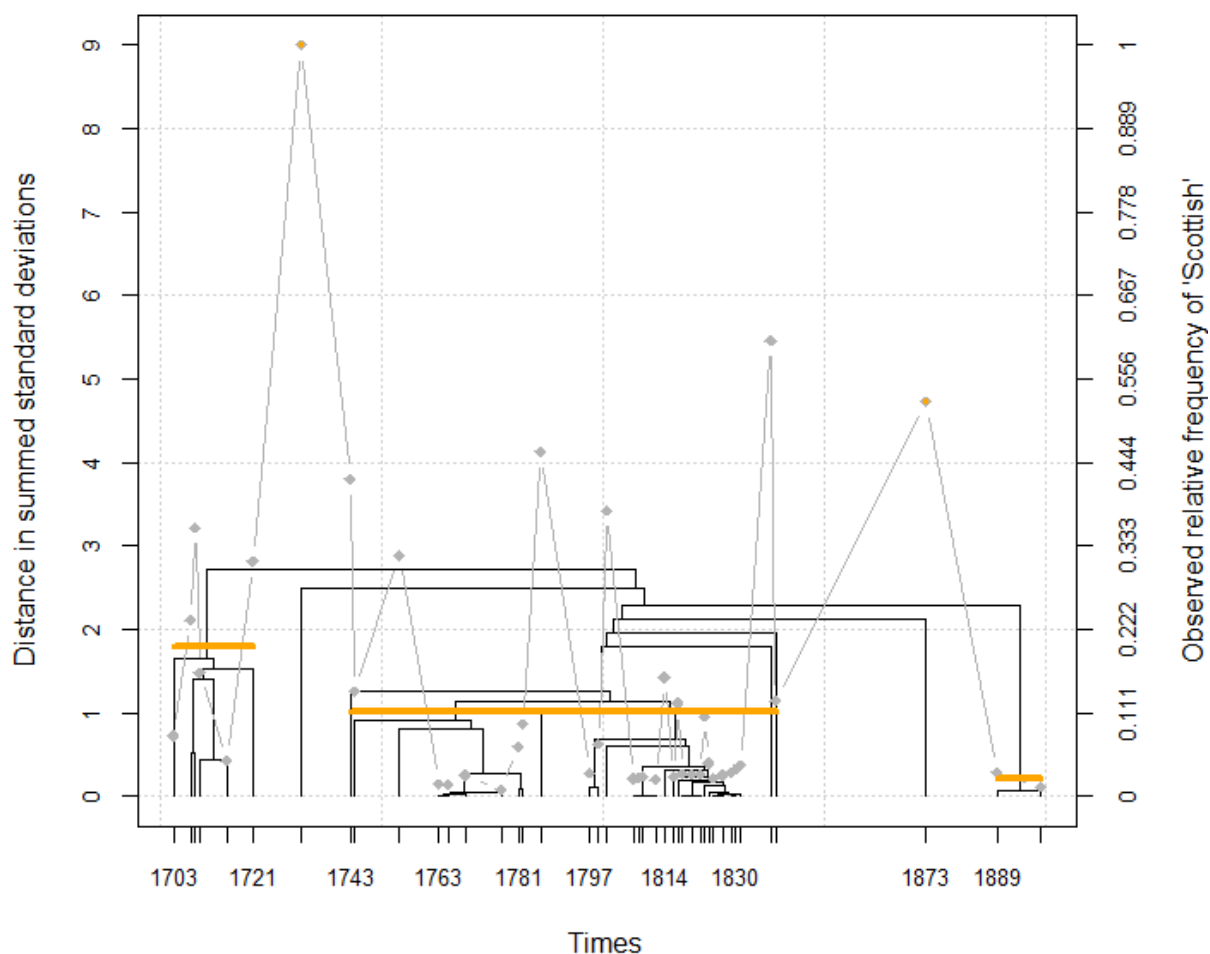


Figure 9: VNC of levels of Scots words for those known to have pro or anti Scots sentiments (5 clusters)

The patterning of the data for this particular set of people is not radically different to the overall trends observed in Figure 8, but we do see higher levels of Scots overall – the mean of the observed relative frequency for the years 1744-1837 is at 0.112, compared to 0.037 for the general Scottish population. It seems those with pronounced political viewpoints did use more Scots in their writings overall. Various genres are represented by this subset of people in the corpus, and so this effect is not the result of Scottish vernacular poetry or imaginative prose alone. Of course, the frequency observed here may have been bolstered by the ‘anti’ camp, yet this still indicates that political affiliation did seem to influence people’s use of Scots – these authors are behaving marginally differently to the general Scottish population. Use of Scots is particularly high in the first cluster,

which spans the years 1703-1721. This period saw the Union of the Parliaments being negotiated and initiated, and three of the four Jacobite Risings (1708, 1715 and 1719) taking place. It is plausible that these major cultural and political events had an influence on the writings of politically-involved individuals active during this time, particularly those in the opposing camp. The Union agreement and its negotiation process was by no means welcomed by all sectors of Scottish society, and various Scotsmen identified with the goals and ideology of the uprisings instead. However, it must be remembered that the timeline presented here is but one section of a change in progress, in which written Scots was declining, and hence we may simply be seeing a snapshot of that decline. Nonetheless, the notably higher frequency of Scottish lexical items in this dataset compared to general Scottish society suggests other extralinguistic factors contributed to this trend.

Again, the years 1744-1837 cluster together. Just as the mean Scots usage among general society appears to have remained stable during this time period, so too the opposing political sentiments seem to have created an equilibrium during the second half of the eighteenth century. The increasing dissatisfaction with the Union, patriotic defiance towards the establishment position and the radical movement seem, if nothing else, to have slowed or momentarily halted the decline of written Scots. The relatively high standard deviations of the merged files across the time frame imply constant variation, rather than a sudden rise in Scots usage during the eighteenth century. This similarly suggests the divergent influences (in particular political loyalties) operating on these authors. The raw data overlaid on the graph also reflects sudden fluctuations between high levels of Scots and almost categorical English use, which again could be attributed in part to the two opposing political viewpoints of the authors. These peaks and drops are consistent throughout this time period, suggesting two different camps of writers were at play; those that used relatively high levels of Scots in their works, and those that used almost no Scots in their writing. It is possible that these two patterns could map onto political affiliation, and this will be explored further in section 5.3.4.

5.3 Statistical Modelling – The Extralinguistic Factors

To answer research questions three to five, I sought to use a number of statistical methodologies to examine the importance and influence of the independent variables (the predictors) upon the variation in the data. The use of linear or logistic mixed-effects regression models is well-established within sociolinguistics in particular, as these provide a powerful and principled way to examine multiple predictors simultaneously (Baayen et al., 2008; Johnson, 2009; Tagliamonte & Baayen, 2012). In recognition of their strength and quantificational accuracy, I initially tried to run a logistic mixed-effects regression model on the data. However, the dataset was extremely imbalanced and unsurprisingly, the mixed effects model failed to converge. A number of different solutions were trialled, none of which were successful¹⁴. Instead, I opted for other statistical models that are better suited to imbalanced data, which so often characterises historical analyses. Thankfully there are a number of statistical models that are able to handle the inconsistencies of historical data whilst still providing an accurate and robust examination of the extralinguistic factors conditioning variation. These included conditional inference trees and random forests. Their properties and their application to my data will now be explained in more detail.

5.3.1 Conditional Inference Trees

Conditional inference trees (or *ctrees* as they are commonly referred to) are non-parametric, tree-structured regression models embedded within a conditional inference framework (Hothorn et al.,

¹⁴ Regrouping predictor levels into larger subsets, excluding non-variable authors, treating time as non-linear by fitting a cubic spline, excluding time altogether as a predictor (given the extremely uneven distribution of the data across time) and examining only the twenty-four politically-active individuals were all trialled. None were successful in getting a model to converge, even with a single predictor the model failed. There was only a very small window of variation for the model to work with, and a number of predictors barely reflected any variation at all. Furthermore, some of the predictors may exhibit collinearity (such as profession and genre, for example). All these properties can be highly problematic for mixed-effects models, which can become severely destabilised by imbalanced data and potential collinearity between predictors. It may simply be that more Scots data is needed in order to balance the data better, or that a more complicated model is required for such a complex dataset. Neither of these options were viable given the timing constraints and amount of extra work they required, with no particular promise that they would be successful in achieving model convergence.

2004: 1). They are similar in nature to regression models in that they are able to statistically examine the relationship between multiple predictors (the social and linguistic factors influencing the variation) and the variable, but they present the interactions in the data as a tree-model instead, (Tagliamonte & Baayen, 2012). To determine the significance of a predictor, ctree uses recursive binary partitioning, which refers to the process whereby the algorithm estimates the likelihood of the value of the response variable (*Scots* or *English* in this case) based on a series of binary questions about the values of the predictor variables (the levels or categories of a predictor). So, for example, it will consider whether splitting the data by *Pro-Union* and *Anti-Union* authors (for the predictor POLITICAL AFFILIATION) will align with the linguistic data, so that one branch has a greater level of *Scots* and the other branch a higher level of *English*. The ctree splits the data by predictor into partitions like this again and again, working its way through all the predictors we choose to include in the model. Each partition is recursively analysed, to test for its level of in-group similarity. The ctree is looking for relatively homogenous data partition – this indicates that either a high level of *Scots* or a high level of *English* tokens is present in that data partition, which in turn suggests that the particular predictor splitting the data is useful in predicting the response.

Each predictor chosen forms a 'node' in the ctree, and each of the partitions form binary 'branches' stemming from this node. Each split is also statistically significant ($p < 0.005$) – in other words the interaction between the predictor and the variable is significant. The predictor that is the most significant in determining the response variable will be selected first and this forms the 'root' of the tree. Branches are constructed off either side, and with each division the ctree tries to create an optimal split. This carries on until further splitting no longer gives us high similarity between the data points, or until the tree has reached the maximum depth (number of levels) specified by the researcher. A test of independence is also carried out between each predictor and response. This indicates how much predictive power is lost if the predictor is removed – again indicating how well the response and predictor variables align. If independence is indicated, the predictor is not useful, and the next predictor is trialled instead. At the terminal nodes of the tree (the predictors lowest in

the tree) the proportions of the variants are depicted as a series of bar graphs, allowing the researcher to quickly and easily see how the different variants are concentrated across interactions with the predictors.

Ctrees can also incorporate random effects into the model, an important consideration given that there is always 'random' variation present in a data set. This can stem from the behaviour of individual authors or variants, and these effects are not repeatable, but rather taken to be representative of a much larger population and language pool (Baayen, 2010). If this is not controlled for, the relationships found in the data hold only for the authors and words examined, and cannot be extrapolated to the wider speech community (Baayen et al., 2008). By incorporating speaker or author as a random effect, statistical models predicting language variation and change are thus able to recognise where a potentially crucial source of the variation comes from, which in turn can prevent them from potentially overpredicting the significance of the predictors included in the model (Baayen, 2010; Baayen et al., 2008; Johnson, 2009; Tagliamonte & Baayen, 2012).

Ctrees are thus able to uncover the myriad of interactions between the variants and the predictor levels, as well as forming a useful visualisation tool to suggest the fine-grained distinctions among the different interactions in the data (Tagliamonte & Baayen, 2012: 164). They are able to provide a more intricate examination of the data than regression models, and are able to handle messy, imbalanced data more easily as they make no assumptions about its distribution. Especially in diachronic change, developments are rarely conveniently linear and often involve complex non-linear relationships between variables (Gries & Hilpert, 2010). Some effects may apply only for a certain window of time, or to a certain subset of the data, but ctrees are able to tap into the complex profile of a variable and the various conditioning factors that may be responsible for its manifestations.

However, ctrees are liable to overfitting, and very sensitive to changes in the data. They are not as quantificationally robust as mixed effects models or random forests. Though they are able to portray

the delicate interplay between different factors and the variable under investigation, they lack the power of other models to fully identify just how accurate a particular predictor is in determining the response. This can be mitigated to an extent by keeping the tree short (allowing the tree to grow to just three or four levels), determining the size of the partition (each partition must contain at least a certain number of observations) and ensuring that the ctree only indicates splits that are statistically significant ($p < 0.005$). An alternative, however, is to keep the tree weak (i.e. allowing them to grow as large as the algorithm decides) but instead grow several hundred trees and take a collective vote from the trees. This is a random forest, and this is explained in more detail below.

5.3.2 Random Forests

Random forests are well suited to historical datasets as they are able to handle widely unbalanced datasets with high multicollinearity, especially given that highly imbalanced cells and correlated factors can be hugely problematic for mixed-effects models, severely destabilising them and forcing the researcher to remove various predictors until convergence is reached (Tagliamonte & Baayen, 2012). They are able to examine the importance of multiple predictors even with a small number of observations (tokens), another feature often characteristic of historical data. This provides the random forest with much greater quantitative power than an individual ctree, and improves the algorithm's ability to examine the relationships of the predictors with the variants under examination, even these are disproportionate within the dataset (i.e. one variant occurs much more commonly than the other). This was a particularly important consideration for the data under investigation here, given that the Scots lexemes only comprised of eight percent of the dataset (and English the remaining ninety two percent), which may explain why the mixed effects models continually failed.

A random forest is essentially a large number of individual ctrees, each of which contain a subset of the data, by randomly sampling without replacement from the standard dataset (observations and predictors). For each tree its training set (the sample) is paired with a test set (the remaining data

not included). The accuracy of a tree's predictions is evaluated by comparing its predictions for the test observations with the actual values observed for the test data (Tagliamonte & Baayen, 2012: 159). This ensures greater accuracy in the predictions being made by the trees, improving the accuracy of the overall result. Using the same process of trial and error as an individual ctree, the trees select which predictor they find to be most important in describing the variation (the predictor that forms the top node of the tree), which they then contribute a vote. The forest collects these votes and ranks the predictors according to how commonly they occur in the top node. This ranking indicates their variable importance – how well each predictor can determine the response variable (*Scots* or *English*). This averaging approach reduces variance and bias, and thus the possibility that outliers seriously interfere with the overall trends in the data. The trees are also decorrelated – each tree is only given a subset of the predictors to evaluate. This prevents one particularly powerful predictor from dominating the dataset entirely, otherwise it would be chosen every time by the trees, and no other predictor would ever stand a chance of making it into the importance measures. The forest is thus able to consider all predictors on an individual basis, and then identify which explains the greatest amount of variation (Tagliamonte & Baayen, 2012).

Taken together, random forests and ctree can provide a comprehensive, systematic examination of diachronic change without the need to rely on generalised mixed effects models, which are liable to failure due to the very nature of historical data. Random forests present a novel type of non-parametric data analysis (Tagliamonte & Baayen, 2012: 136), that provide the researcher with an overall view of the nature of the dataset and the differing weights of the predictors, regardless of the number of predictors or high levels of empty cells. The ctree adds to this perspective by uncovering the specific interactions of these predictors, and how the data is stratified across the various interactions between the response variable and the predictor levels, before presenting this in a visual, easily-identifiable format.

The VNC analysis above has already answered research questions one and two, suggesting how the frequency of Scots lexis patterned over time for the general and politically-active Scottish society. To answer the remaining research questions - which socio-historical factors were most important in influencing the frequency of Scots lexis in general society and in the politically-active sector of Scottish society, and whether authors with political sentiments reflect different frequencies of Scots lexis along pro- or anti-Union lines – ctree and random forests were utilised. The socio-historical factors (the predictors) included for analysis, along with their predictor levels, are presented in Table 3 below.

Table 3: Predictors and predictor levels used for random forest and ctree modelling of general literate Scottish society

Predictor	Predictor Levels
Genre	Administrative Prose Political - Prose Imaginative Prose Instructional Prose Orthoepist Political – Creative Journalism Personal Writing Expository Prose Verse/Drama Instructional Prose Correspondence – Political Religious Prose
Profession	Politician Author

	Poet Legal Professional Orthoepist Other
Education	Boarding School Parish School University Unknown Apprenticeship Secondary School
Birthplace	Glasgow Edinburgh Scotland_Other England France Aberdeen Unknown
Place Published	England Glasgow Unknown Edinburgh Scotland_Other America Australia Europe

	Ireland
Pro_or_Anti	Pro Anti Unknown
Gender	Male Female
Year	1700-1740 (VNC1) 1740-1837 (VNC2) 1837-1860 (VNC3)

The remaining predictors not included in the statistical models include the following: FATHER’S PLACE OF BIRTH, MOTHER’S PLACE OF BIRTH, TITLE (Author’s Title), RELIGIOUS AFFILIATION, PLACES RESIDENT (countries where the author resided during their life), PUBLISHER, YEAR OF BIRTH and LANGUAGES SPOKEN (languages the author was fluent in other than English). These were not included as there was very little information available for these categories (for example, MOTHER’S PLACE OF BIRTH consisted of 82% *Unknown*). This complicates a statistical model’s ability to find a robust interaction between the remaining values, especially as it will treat *Unknown* as a category and therefore could indicate an effect where there is none.

5.3.3 General Scottish Society

5.3.3.1 Random Forest

To answer the third research question – which sociolinguistic factors were most important in influencing the frequency of Scots lexis in general Scottish society as a whole - it is useful to grow a random forest. This is able to assess the variable importance of the different predictors. Accordingly, a random forest was grown using the **ranger** (Wright and Ziegler, 2015) package in the open-source,

user-extendable, statistical platform R (R Core Team, 2013)¹⁵. All predictors were included (see Table 3), and the data from the entire corpus was used, with Scots words vs English words as the dependent variable. AUTHOR and TEXT were included as random effects in the model. The politicians and politically-active authors were included in this data, as to remove them would be removing one of the groups that made up the general literate Scottish society, creating a somewhat artificial composition of the literate sector active during the eighteenth century. Furthermore, as they comprised of just twenty-four out of a total of 134 authors, any potential effect for political actors will not dominate the general dataset. Both YEAR OF PUBLICATION and YEAR OF BIRTH were trialled, but these seemed to be behaving almost identically. They both occupied the same slot in the importance measure rankings, suggesting they explain the variation in the data equally well. Given that there was more consistent data for YEAR OF PUBLICATION than YEAR OF BIRTH, which was often unknown, YEAR OF PUBLICATION was chosen as the numeric measure to be included (relabelled to YEAR). This then enabled YEAR to be recoded according to the main VNC clusters identified in the analysis earlier. The rankings of the predictors are shown in Figure 10 below.

¹⁵ The seed was set to 89788 and the importance measure was set to impurity. Impurity refers to the Gini index. This is a measure of node purity – a small value indicates that a node in a conditional inference tree contains predominantly observations from a single class (Scots, in this case). Thus, the smaller the value for impurity, the greater the number of observations of a particular variant for that predictor, and hence the stronger the predictor is in determining the variation observed (Witten et al., 2013)

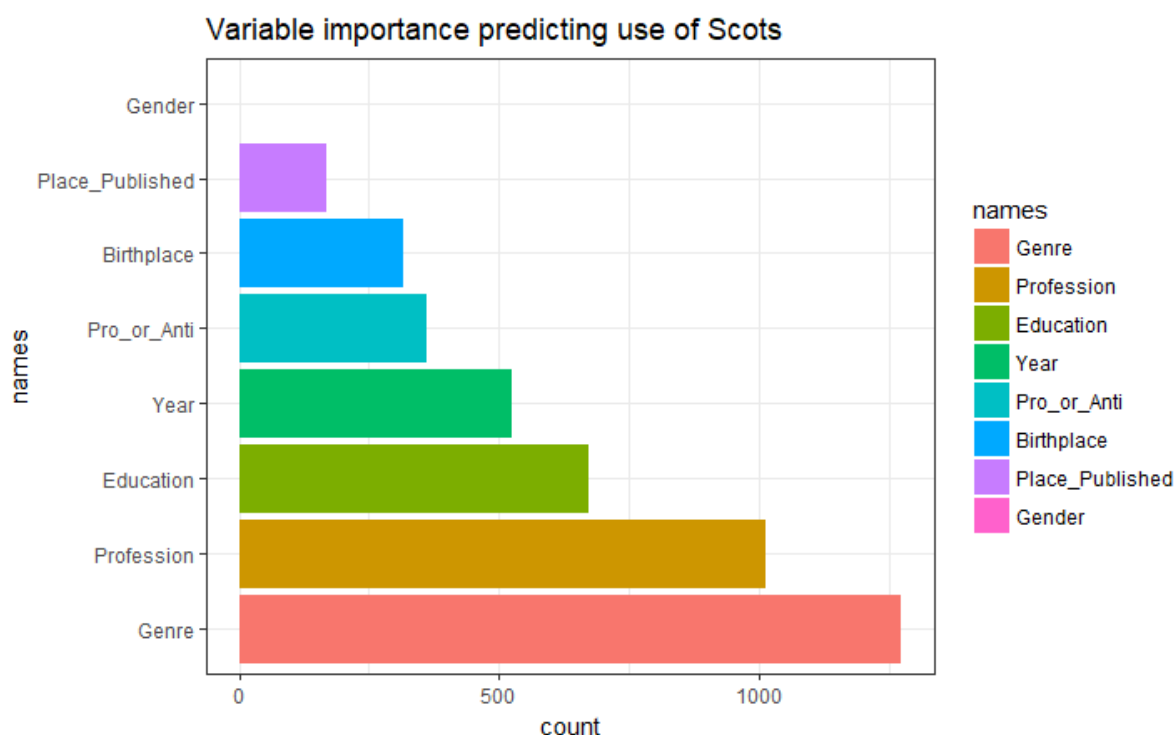


Figure 10: Variable Importance measures from random forest for general literate Scottish society

Mean Gini Decrease

The different predictors are listed along the y-axis. The further these stretch along the x axis, the stronger they are as a predictor conditioning the use of Scots or English. The Mean Gini Decrease refers to the Gini index. This is a measure of node purity – a small value indicates that a node in a conditional inference tree contains predominantly observations from a single class (Scots, in this case). Thus, the greater the decrease in the Gini index, the greater the number of observations of a particular variant for that predictor, and hence the stronger the predictor is in determining the variation observed.

Examining Figure 10, it is clear that GENRE stretches furthest along the x axis, suggesting it explains most of the variance in the data. It seems the genre of the text being produced played the greatest role in determining authors' choice to use more or less Scots lexis. This is not a surprising observation – there were different expectations and goals surrounding different genres. Texts aimed

at the wider public were frequently anglicised to a higher extent, as these sought to reach not just the local clients but the readership beyond Scotland's borders (van Eyndhoven & Clark, forthcoming). The audience for Scottish texts had been expanded by the Unions of 1603 and 1707, which enabled not only greater dissemination, but also the profitability of printing as a process, given that Scotland at just one million people was not sufficient to finance the expense of printing (Graham, 1908; Clive, 1970). Furthermore, publishing houses and printers in Scotland modelled their practices on the print culture that originated from England and was structured by English models (Harris, 2005a; Pentland, 2011), and the default language of print was English (Millar, 2013).

Yet creative works, which were similarly intended for a wider audience, could be expected to exhibit higher levels of Scots some of the time, depending on the creative (and potentially patriotic) goals of their authors (Buffoni, 1992: 127; Dossena, 2005: 96; Smith, 2007; Corbett, 2013). Administrative texts were also likely to contain higher levels of Scots lexical items, not for creative or patriotic reasons, but as part of their highly codified, formalised nature. Being inherently tied to Scotland's independent legal and religious institutions, these texts often preserved Scots terms and expressions far longer than other texts (Bugaj, 2005, 2013; Kopaczyk, 2012, 2013; van Eyndhoven & Clark, forthcoming). The anglicising pressure was strong for most authors operating during the eighteenth century, but the diverse goals and expectations surrounding different genres arguably led to varying levels of Scots and English across different texts, and this is reflected in the findings of the random forest.

GENRE is followed by PROFESSION in the variable importance measures as the second most important predictor conditioning the use of Scots words. It seems that, following literary constraints, societal pressures and institutional expectations formed the next strongest pressure operating on authors during the eighteenth century. Of course, there may be some correlation between GENRE and PROFESSION – the poets in POLITECS for example are often represented by poetical works, in which they may have used a greater number of Scots words, which the genre allowed and even stimulated.

Yet this is also a benefit of the random forest; it can consider all predictors in their own right, regardless of correlations and then identify which is superior in explaining the variation (Tagliamonte & Baayen, 2012). In this instance, the random forest is still able to show that GENRE is more important than PROFESSION, even if there are significant overlaps.

Despite the correlations, there are also texts in the corpus that do not necessarily correspond to occupation. For example, the correspondence in POLITECS is written by authors that include lawyers, clergymen, academics, weavers, poets and military men, to name but a few. Thus, it seems that not just GENRE, but also the position members of the literate Scottish community held influenced their levels of Scots or English words. Professions clearly came with a whole host of attached attributes, such as level of education, social rank and degree of contact with English or Scots – military leaders or the governing elite would have had far greater exposure to written English on a daily basis than a creative author for instance. In the case of orthoepists (language commentators) such figures were both highly conscious of their language use and had a public presence built upon their profession, thus dictating the desire to be extra vigilant in their language use (Aitken, 1979; Murison, 1979; Smith, 1970; Caie, 2007). The position of PROFESSION as second-most important is therefore not all too surprising, given the time period in question.

EDUCATION is ranked third most important in the random forest, and again there are certain correlations with some of the other factors, such as PROFESSION. Yet it also speaks to the standardising influences of educational institutions across Britain during the eighteenth century. In particular, boarding schools could be expected to have anglicised the writings of Scotsmen. Boarding schools were often located in England or modelled on English institutions. Children attending these institutions would thus have passed through the educational system writing purely in English, without much if any exposure to written Scots (although of course they may have come into contact with spoken Scots). Indeed, the elite often sent their sons to boarding schools in England in order to learn to speak and write in English, this practice becoming commonplace in the latter half of the

eighteenth century (Aitken, 1985). Orthoepists similarly sent their children to English schools, in the hope they would not acquire Scottish features in their speech, but instead obtain a perfect command of English (Jones, 1995). Local parish schools on the other hand were much less likely to be quite so English-focussed, and children attending these schools are likely to have come into contact with written Scots more regularly than their England-based boarding school peers.

Scholars who attended university would have experienced different pressures again, considering the contact they no doubt had with works issuing both from England and wider continental Europe, as well as the expectations surrounding the language of academia. English universities were still being conducted in Latin or French when Scottish scholars began to publish some works in the vernacular during the sixteenth century, especially when these were intended for a wider audience (Bugaj, 2004a). Yet the influence of Latin did begin to break down in the seventeenth century across Britain, and by the end of the eighteenth century works from a variety of fields were being published in the vernacular in increasing numbers (Gordin, 2015). For example, Galileo Galilei's publications were initially in Latin, but his later publications (such as *Opticks* of 1704) was first published in English (Gordin, 2015). The language of academia was thus becoming a mixture of Latin and English. This break-down coincided with the Union of the Crowns (1603) and by the time of the Union of the Parliaments (1707) it was well underway, which no doubt had important ramifications on scholars seeking to export their research to the now much-wider, English-based academic community (van Eyndhoven & Clark, forthcoming).

Scotsmen attending English universities would have developed their academic writing skills in Latin or English, which would have important influences on their future publications, and even those passing through Scottish universities would have experienced increasing pressure to anglicise their work. The Scottish Enlightenment in particular saw a deep insecurity among intellectuals towards their native Scots speech (Clive, 1970; McClure, 1994: 40; Bono, 1989; Dossena, 2002; 2011), which in the very least would have discouraged their Scots writing skills. Those attending university were

also mostly the elite, who had particular social goals attached to their career that demanded the use of English, (Smith, 1970; Aitken, 1979; Jones, 1995; Corbett, 2013; Cruickshank, 2017). Accordingly, the position of EDUCATION as the third more important predictor conditioning authors use of Scots during the eighteenth century can be explained if we consider the nature of the different educational institutions and both the pressures and opportunities they presented to their attendees.

The random forest provides a statistically robust method for modelling diachronic data by the numerous factors influencing the variation contained within it, suggesting which factor is most important in predicting the response (*Scots* or *English*) overall, and where the other predictors rank relative to their variable importance. This gives us a nice overview of which factors operated on general, literate Scottish society, suggesting that GENRE, PROFESSION and EDUCATION were particularly important in determining their choice to use more or less Scots. What we cannot see from the importance measures, is the direction of the influence; we cannot see which genres for example encouraged the use of Scots, and which the use of English. Thus, while the above explanations perhaps explain why these factors have been identified as most important in driving the change, it is as yet unclear whether language use actually follows along these lines.

5.3.3.2 General Scottish Society – Ctree

In order to see interactions between the variants and the predictor levels, it is helpful to grow a single conditional inference tree. This is able to suggest not just the most important predictor, but identifies how the variable is conditioned by different combinations of predictors to form data subsets. The strongest predictor will form the top node – this is the first and most important predictor determining whether a choice was *English* or *Scots*. The lower branches identify predictors with secondary and lower level importance; these will apply to a particular subset of the data.

Accordingly, ctree was grown in R (R Core Team, 2013) using the `ctree()` function of the **partykit** (Hothorn et al., 2010) package¹⁶. The dependent variable was SCOTS_ENGLISH, with the same

¹⁶ The seed was set to 1234.

predictors included as in the random forest (see Table 3). YEAR was included as a raw value this time, in order to investigate whether the ctree could uncover more fine-grained interactions between YEAR and the response variable, or whether the interactions would instead match the VNC analysis. AUTHOR and TEXT were included as random effects in the model. The level of significance each split in the tree needs to reach before this split is made (the *mincriterion*) was set to 0.099 (so $p < 0.001$), the minimum number of observations (number of *Scots* or *English* tokens) required for each branch in the tree (the *minbucket*) was set to 200, and the tree was pruned to five levels deep – so it could split up to five times on a single route, but not more than this (the *depth level*). The *minbucket* was set quite high due to the large number of observations in the data (777,438 *English* or *Scots* tokens), while the *depth level* was set relatively deep due to the high number of predictors included. The output of this ctree is shown in Figure 11 below.

Figure 11: ctree showing proportions of Scots across general literate society with all predictors included

The ctree indicates some interesting interactions between the predictor levels and the dependent variable, and, as in the random forest output, indicates that GENRE is the most important predictor conditioning the variation. GENRE forms the top node (or ‘root’) of the tree, suggesting that it forms the strongest relationship with the dependent variable. However, the ranking of predictors lower in the tree diverges slightly from that of the random forest. This is caused partially by the nature of a ctree itself – though ctree are well suited to identifying interactions in the data, they have less quantitative power than bagging ensemble algorithms such as random forests. Thus, though ctree can capture the complex data structure of a variable, they are less well suited to making judgements on the importance of the predictors determining that structure. The tree model suggests for instance, that PLACE PUBLISHED and PROFESSION are on par in terms of importance for their respective subsets of the data. It is unclear which is ranked higher by the tree, and the positioning of PLACE PUBLISHED in this model could for instance be related to the subsetting of the data. Its importance may be related largely to the creative portion of the corpus, whilst it had relatively less significance for all the non-creative texts. Overall the rankings of the predictors higher up in the tree do roughly correlate with the random forest, but their exact importance isn’t entirely clear.

The true strength of the ctree, however, is its ability to uncover the strength and direction of the interactions between the variable and the levels with independent predictors. The root of the tree is conditioned by GENRE, and it demonstrates a clean split between the ‘creative’ genres on the left and the ‘professional’ genres on the right. The left side includes *verse/drama* – these were plays and poems; *imaginative prose* – these were mostly novels; and *political-creative* texts – these were satirical poems and ballads. There is one outlier in this group; the *orthoepist* works. While the first three genres all stem from a creative or artistic sphere, it is more difficult to see where *orthoepist* works fit into the picture. Yet although they are not inherently imaginative writings, these works often contained many Scots lexical items to demonstrate the differences between “incorrect” Scots and “correct” English usage in grammar, sentence structure and pronunciation, which can account for the higher levels of Scots observed in these writings.

Orthoepist aside, Figure 11 clearly demonstrates the vernacular revival or backlash; the creative genres on the left contain the higher levels of Scots, indicating that Scots both continued, and perhaps even increased in use in these literary fields. Though figures such as Robert Burns and Robert Fergusson are icons of this movement, it is clear from the corpus that their contemporaries similarly used higher levels of Scots for their works, than their non-creative peers. This also demonstrates the divide that was by now well entrenched into literate Scottish society; Scots was acceptable in vernacular literature and creative works as part and parcel of the traditional, rustic and historic ideals associated with it (Dossena, 2002, 2005; Millar, 2013), whilst English, as the language of profession and propriety was reserved for most remaining types of prose (Murison, 1979; Corbett, 2013). Regardless of personal sentiments, authors within the creative sphere had more liberty to use Scots without the suppressive constraints facing authors trying to publish more serious types of prose to a general audience.

Moving further down the creative side of the tree, the next most important predictor for this collection of texts is PLACE PUBLISHED. The tree splits into *Edinburgh/Scotland_Other* on the right and *England/Glasgow/Unknown* on the left, and there appears to be a slightly higher concentration of Scots under the *Edinburgh* branch. This is an interesting observation, given that Scottish Standard English was emerging out of Edinburgh towards the end of the eighteenth century. This standard was based upon the language of the legal, clerical and educated circles of Edinburgh, and towards the end of the eighteenth century there was increasing recognition that this standard could be equally acceptable in high society (Aitken, 1979; Dossena, 2011; Smith, 1996, 2007; Corbett, 2013; Jones, 1995). We may be seeing some of this effect here, although this possibility is vague at best given this effect for *Edinburgh* is only observable on the creative side of the graph.

If we examine the non-creative side of the tree, the data is next split by PROFESSION which divides into two categories; *Politician* and everything else. This indicates that politicians were behaving differently to other members of literate Scottish society, at least within the non-creative sphere.

Their position as administrators of Scotland's local state of affairs and involvement with political matters did seem to have an effect on their written language usage. Furthermore, if we examine the *Politician* branch overall, we see that this subset contains the largest proportion of Scots within the non-creative division of the tree, suggesting slightly higher use of Scots by politicians, even when publishing serious forms of prose. Although the level of Scots is clearly not as concentrated as the creative genres, the fact that there is a visible concentration here is still noteworthy, given the constraints on those publishing outside the creative sphere.

This general ctree thus provides a nice overview of the trends in the data, as well as indicating where the data clusters and which factor levels group together. The ability of ctrees to visually demonstrate which levels within a predictor align provides another benefit; they allow us to conservatively and accurately group together different levels within each predictor that are behaving similarly. This can provide more quantitative power to help to uncover robust relationships in the data. Although this does remove some of the finer details in our data, it also provides us with greater oversight into the overarching influences acting upon the variable within different factors. Having established the clustering of levels in Figure 11, predictor levels were now reclassified for the remaining ctrees grown from the data. GENRE was grouped into *creative* and *non-creative* texts using the split identified above (relabelled as CREATIVE), and BIRTHPLACE was condensed to three main categories; *Edinburgh, Scotland_Other* and *Other*. To see a full list of changes made see Appendix 3.

5.3.3.3 General Scottish Society – Political Texts in Ctree

In order to explore the effect of political change on language use, I also sought to observe whether political material might induce the authors to use higher concentrations of Scots than other genres. Authors producing works that were political in nature may have consciously or subconsciously been affected in their linguistic choices by the topic of their discourse. To further uncover whether the sociohistorical factors influencing the frequency of Scots lexis in general society (RQ. 3) were related

to the political changes taking place, texts with a political focus needed to be examined. Texts were thus grouped into *political* and *non-political* texts (labelled POLITICAL). The *political/non-political* distinction was made for both texts from the political component of POLITECS and the general component (the subset of the CMSW, some texts of which were political in nature). Texts were defined as *political* if they discussed opinions related to the Union, the Jacobite Uprisings, Napoleonic wars (which some Scots supported, and some did not), reform of the existing political structure, socio-economic and political relations between Scotland and England, or if they were satires, the product of Scottish radical societies, or political poetry. Though this may have missed a number of texts that are to some extent political in nature, a conservative approach is more desirable to avoid mislabelling texts. As a result, only those with an overtly political agenda were included.

PROFESSION was also regrouped into six categories according to which field the author primarily belonged to; *religious/legal*, *politician*, *academic*, *author_creative*, *author_noncreative* and *poet*. *Author_creative* and *poet* were created as two separate categories as there were high numbers of each (*author_creative* includes novelists, playwrights and songwriters), thus to combine these into one category absorbed a large portion of the writers in the dataset, vastly dominating the PROFESSION. Furthermore, I was interested to see whether poets might behave any differently from creative authors, given that poets could and did use their works to make veiled political comments (Dossena, 2005; Smith, 2007), which might in turn affect their use of Scots. I also excluded all authors who did not vary between Scots and English. Some authors used exclusively English, which although interesting, does not provide much information on what was driving their linguistic choices, as they were not making a choice between Scots and English.

Using this dataset that contained the newly categorised predictors, and variable authors only, another ctree was grown, using function `ctree()` in the **partykit** (Hothorn et al., 2010) package in R. The independent variables included CREATIVE (whether the text was *creative* or *non-creative*),

POLITICAL (whether the text was *political* or *non-political*), BIRTHPLACE, PROFESSION and YEAR (again raw values were used), the dependent variable was SCOTS_ENGLISH and AUTHOR and TEXT were included as random effects.¹⁷

¹⁷ The mincriterion was set to 0.95, the minbucket was set to 200, and the depth level set to 5. The seed was set to 1234.

Figure 12: ctree showing proportions of Scots for general literate Scottish society with genre recoded to include Political texts

It is clear from Figure 12 that the *creative/non-creative* divide is still the most important predictor determining use of Scots, despite the restructuring of the predictor levels, and the exclusion of non-variable authors – CREATIVE is positioned in the top node. This once more confirms the presence of the vernacular backlash.

Interestingly, the next predictor influencing the creative genres is POLITICAL. This indicates that it is a relatively important predictor for general, literate Scottish society, and more interestingly still; we can observe higher concentrations of Scots within texts that are *political*. This implies that the greatest levels of Scots across eighteenth century authors can be observed in texts that are both creative and political. Such texts would have included satires, radical plays or novels, satirical songs/ballads and political poetry. We know that some poets such as Robert Fergusson utilised the medium of poetry to make veiled political comments or overt patriotic remarks (Dossena, 2005, 2012; Smith, 2007), but also as an outlet for creative expression in Scots that was not available in other genres (MacDonald, 2011). There is a tangible link between patriotism, an increasing political awakening and the vernacular revival, thus it is perhaps not unsurprising that we see the highest levels of Scots where these fields overlap. Creative literature already exhibits higher levels of Scots than other genres, as we have seen in Figure 11 and can see again in Figure 12, but it seems these levels were further elevated when used in political contexts. It is plausible that the influence of politics stretched to the cultural sphere, where ideas of nationalism and identity perhaps encouraged further use of Scots.

The effect of political texts on use of Scots is thus fascinating and promising, but it is also interesting to look briefly at the non-creative side of the tree, to observe any potential differences. PROFESSION (as in Figure 11) again affects the non-creative side, and once more *politician* branches off from the other occupations, although now they are grouped with *religious/legal* professionals (which contains all authors working in religious or legal fields in some form or another; including advocates, treasurers, clergymen and bishops). This branch similarly contains the highest proportion of Scots in

the *non-creative* branch of the tree, suggesting that politicians are still behaving differently in their language use from the general (though now variable) society, although they are joined by authors from the legal/religious fields. This is perhaps not entirely unsurprising; there is frequent overlap between the three fields and many politicians included in our corpus were also lawyers during their lifetimes, or involved in legal occupations. It is not unfeasible that the independent religious and legal systems of Scotland would have interacted with the political circle on a regular basis in order to be maintained and managed.

The relatively high use of Scots for those in *religious/legal* fields may also be due to their interaction with written Scots on a daily basis, and the emergence of Scottish Standard English. Recall that various Scots lexical items were maintained by Scotland's independent religious, educational and legal institutions, which remained intact after the Union of the Parliaments. These were words or lexical bundles (Kopaczyk, 2012, 2013) that were specific to these fields, and for which an English equivalent did not exist. This is one explanation for why certain Scots lexical items persisted within certain religious or legal registers long after their use had died out of general usage (Bugaj, 2005; Kopaczyk, 2012, 2013). Authors working with these texts were thus exposed to written Scots more frequently than many of their contemporaries. Furthermore, there are indications that a Scottish Standard was emerging out of the language of these professions towards the end of the eighteenth century (Aitken, 1979; Dossena, 2011; Smith, 1996, 2007; Corbett, 2013; Jones, 1995). It may be that the heightened levels of Scots in Figure 12 reflect the beginnings of this movement; those operating in such fields use more Scots in their writings, both as part of the profession, but also because its use was not stigmatised in the same way that other vernacular forms were. Thus, even on the creative side of the tree the *religious/legal* professionals branch off as a separate group, together with the *poets*, again exhibiting higher levels of Scots. It is plausible that a combined influence of the creative sphere and the demands of the profession induced higher levels of Scots in their writings, though this demands greater investigation before such claims can be made with confidence.

5.3.3.4 Results Summary

In terms of general literate Scottish society, both the random forest and the ctrees have demonstrated that a number of predictors in particular affected the choice to use more or less Scots. GENRE is clearly the most significant, with a clear split by *creative* and *non-creative* works (minus the orthoepists) demonstrated by the ctrees. Those producing work within the creative sphere reflect higher levels of Scots than their contemporaries writing in more serious, non-creative prose types.

Following GENRE, PROFESSION was found to be next most important, though Figures 11 and 12 have demonstrated that this applied mostly to non-creative prose types. Politicians were found to use more Scots compared to other professions, while a further analysis that examined GENRE more closely (Figure 12), found that POLITICAL texts also reflected proportionally higher levels of Scots. It appears that creative texts had higher levels of Scots in general, but these were further elevated when used in works that were political in tone. It may be that ideas of identity and nationalism were strong influences within this subfield.

Through the reorganisation of the predictor levels, Figure 12 also indicated that politicians, the clergy and legal experts grouped together as the sector of society using the highest levels of Scots. For those operating within religious and legal fields, this can be explained in part by the use of certain codified expressions that retained Scots forms, and the rise of a Scottish English standard emerging from their institutional circles. The higher levels of Scots in politicians on the other hand, cannot be explained solely by these factors, given that many operated in England or worked with native English speakers for considerable lengths of time. There must be an additional reason, and it is this group, and their politically-active peers, who form the next part of the investigation.

5.3.4 Politically-Active Scottish Society

5.3.4.1 *Random Forest*

Having examined what was happening in the cross-section of Scottish society represented by the corpus, the final step was to explore how the politically-active authors in the corpus were behaving. I sought both to compare the factors affecting their Scots usage to those affecting general society, and to determine whether their Scots usage was split along Unionist lines. A subset of authors with identified political sentiments was once again created from the data. This was the same subset created for the VNC analysis (see Figure 9) - it contained the authors that constitute the political component of POLITECS, and a number of authors from the general component, who were identified as political during the research phase. This amounted to twenty-four authors in total.

The first step was once again to grow a random forest. Using the subset of data that contained only the authors with known political sentiments, a random forest was grown using the **ranger** (Wright & Ziegler, 2015) package in R (R Core Team, 2013), with the following predictors included: CREATIVE, POLITICAL, PRO/ANTI, BIRTHPLACE, PLACE PUBLISHED, EDUCATION, PROFESSION and YEAR¹⁸. To aid the random forest in identifying the strength of different predictors in the data; the reclassified categories were used for GENRE (CREATIVE and POLITICAL), the clusters identified by the VNC analysis (see Figure 9) were used for YEAR¹⁹, and the further-collapsed predictors, based off the behaviour of the previous ctrees (see Appendix 3). Multiple predictor divisions only reduce the quantitative power of the forest, thus by binning these together greater statistical accuracy can be achieved, particularly when examining smaller subsets of data.

Once the random forest had been grown, the index of concordance (C) was extracted from the forest. This is a measure of accuracy; it tells us how reliable the model is by validating the predictive

¹⁸ The seed was set to 89788 and the importance measure was set to impurity.

¹⁹ This also followed the same methodological practice applied to the random forest grown for general Scottish society. A second random forest was also grown that included YEAR as a raw value, to validate this approach. This slightly decreased the importance of YEAR as a predictor, although the difference was minor. The VNC clusters were thus chosen for the random forest modelling.

ability of a model. $C = 0.92$ for this random forest, indicating that the model was 92% accurate, which seemed promising at first glance. However, when the confusion matrix was extracted from the forest, it indicated that the algorithm was not predicting the minority class (*Scots*), but rather the majority class – the *English* tokens in this case. As a result, the model was predicting the occurrence of *English* tokens very accurately, but was also predicting a large number of *Scots* tokens incorrectly as *English*. The *English* proportion of the dataset, unsurprisingly, was 92%, vastly dominating the response variable. These high levels of *English* are not improbable, given the overall anglicisation trends of the eighteenth century. The result of such trends, unfortunately, is that this creates vastly imbalanced proportions of *Scots* to *English* in the data. The minority class (*Scots*) is overwhelmed by the majority class (*English*), which makes it more difficult for the machine learning algorithm to identify relationships between predictors and the minority class. Instead, the algorithm learns from the composition of the data that it will achieve high accuracy if it predicts the majority class. This approach does produce high accuracy, but this is only reflecting the underlying class distribution rather than the true relationships in the data (Brownlee, 2015).

Although random forests are well suited to exploring imbalanced data, they can handle such datasets only up to a certain degree before they begin to struggle, especially with a 92-8 percent distribution of the two classes. This effect becomes more apparent when we are dealing with just twenty-four authors. However, this can be reduced by subsampling the data. Subsampling is a technique to reduce imbalance in datasets, and there are a number of tactics to achieve this, including upsampling, downsampling, and various combinations of these two methods.

Downsampling (or undersampling as it is also known) was chosen as the best method to reduce the imbalance, given that we were working with a very large dataset (although we only had a relatively small number of authors, we still had 445,789 *Scots* and *English* tokens in total)²⁰. Downsampling

²⁰ This component of the data analysis could not have been achieved without the advice, support and guidance of Dr Vica Papp, whose invaluable statistics knowledge greatly directed this data exploration process, and improved the accuracy of these results.

deletes instances from the over-represented class (Brownlee, 2015), randomly subsetting all the classes (*Scots* and *English*, in this case) in the training set (the subset of the data contained within each ctree that makes up the random forest), so that each class frequency matches the least prevalent class (Kuhn, 2008). Thus, the *English* class was subset to eight percent to match the *Scots* class. The result is that only 16% of the total training set is used to fit the model. This does remove a large chunk of the data from the dataset, as well as some explanatory power because the predictors are weakened. However, downsampling effectively levels the playing field, as the minority class is no longer being overwhelmed by the majority, and it does so in a carefully balanced and measured manner. By randomly subsetting the *English* tokens, this prevents an entire section of the dataset from being deleted (for example removing all academics in the dataset), which could disguise a potential effect. This also helps to maintain the same proportion of *Scots* to *English* across the various predictors. As a result, the trends and relationships in the data are maintained, while some of the bulk is removed. In such instances downsampling can effectively remove the need to collect more data – rather than needing to obtain more instances of the minority class to balance the sample, downsampling can achieve this in a stratified and randomised manner.

5.3.4.1.2 Random Forest - Downsampling

The package **caret** (Kuhn, 2008) in R (R Core Team, 2013) enables the user to subsample random forests, using the function `traincontrol()` (Kuhn, 2008). Accordingly, a second, downsampled random forest was grown in R from the same subset of data and the same predictors specified above (section 5.3.4.1), using the **caret** (Kuhn, 2008) and **ranger** (Wright and Ziegler, 2015) packages and the function `traincontrol()`, with `sampling` specified as "down". The downsampled random forest was then fit using the **randomForest** (Liaw and Wiener, 2002) package in R (R Core Team, 2013). The number of trees was set to 500 (the number of ctree that are included in the random forest), the *node size* set to ten (the minimum size of terminal nodes – so how many observations

(*Scots* or *English* tokens) must be included in the terminal node)²¹, and *importance* measure set to `TRUE` (the importance of the predictors must be accessed).

Once the random forest had been grown, the index of concordance and the confusion matrix were once again extracted. The concordance index was reduced from the original ($C = 0.78$), but the confusion matrix indicated that the model was largely predicting *Scots* this time, rather than *English*. Despite the slight loss in accuracy overall, and the reduction in data, the new model is much better at predicting when the authors chose to use *Scots*, which is ultimately pertinent to this investigation. The variable importance measures were then extracted from the random forest and plotted on a bar graph. These are shown in Figure 13 below.

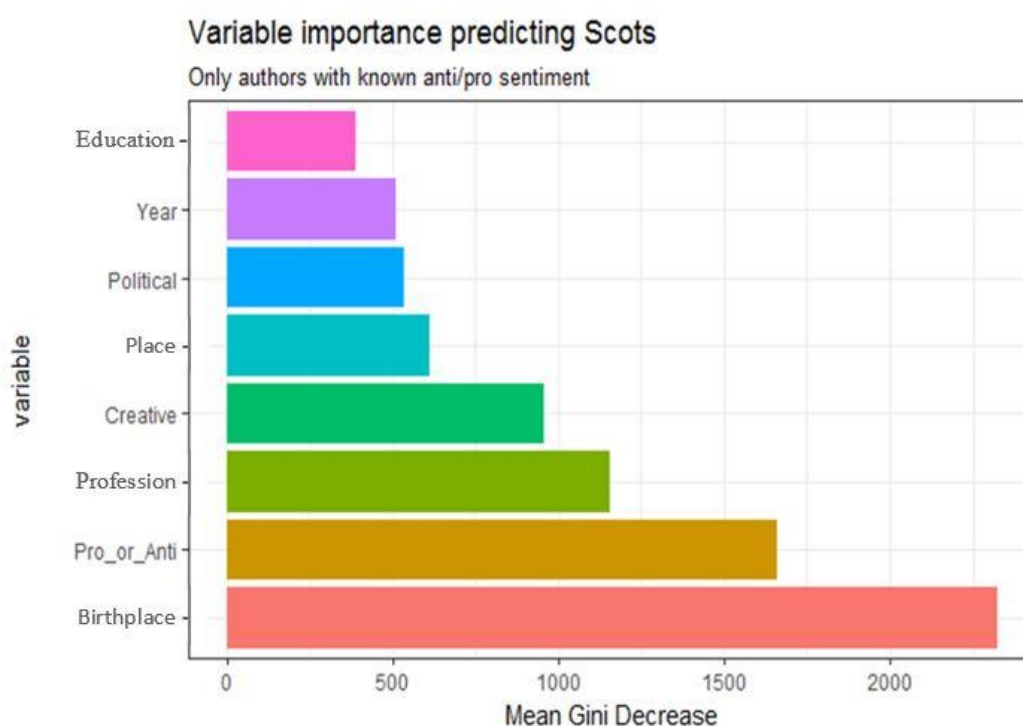


Figure 13: variable importance of factors predicting use of *Scots* for politically-active authors

²¹ Although the default setting for `nodesize` is 1, as this provides the best accuracy, in large data sets a larger node size can be set as this requires less memory and CPU usage, and greatly increases the processing time. For large data sets this normally results in only a small loss of accuracy (Breiman & Cutler, 2018).

Examining Figure 13, the first thing to observe is the strength of an unexpected predictor; birthplace. The random forest suggests that BIRTHPLACE is the most important predictor for these politically-active authors, it seems the immediate surroundings of these individuals largely conditioned their choice to use more Scots or English in their writings. If we consider the nature of the dataset this is perhaps not surprising. Given that the sample comprised of just twenty-four authors, BIRTHPLACE was accordingly grouped into just three categories; *Scotland*, *England* and *Other*. Those born in England are much less likely to have acquired Scots in the writings to the same extent as their Scottish contemporaries, even if they moved to Scotland later in life. Creative authors born in Scotland, for example, could make use of their mother tongue and their complex knowledge of the intricacies of the linguistic spectrum; knowledge that might have been simply unavailable to those born outside of Scotland.

This result could also be partially driven by an interaction between BIRTHPLACE and PRO_OR_ANTI, however. It is not unfeasible to speculate that those who were born and raised in England for example might tend to use more English in their writings and also actively support the Union, whilst the reverse could be true for many authors born in Scotland. Unfortunately, there was not the time to explore how the individual authors were behaving in the corpus, and to compare their positioning along the Scots-English spectrum relative to their birthplace, occupation and political affiliation. As a result, all that can be presented here is the results of the authors collectively. However, this certainly warrants investigation in the future.

What is particularly interesting to note in Figure 13 is the presence of political affiliation (PRO_OR_ANTI) as the next most powerful predictor. This indicates that political affiliation was highly influential in the choice to use more or less Scots for these authors. It may seem self-fulfilling that political affiliation is a highly important predictor conditioning the language choices of individuals who are politically motivated, but how the data is subset does not necessarily determine the lines that language use will follow. There were many social and literary considerations that eighteenth-

century authors had to consider, including the ongoing anglicisation, the pressures of the press, the expectations of elite society and the repression of radical activity. Yet despite these constraints, it appears that the political sentiments held by these authors was a very real consideration, whether consciously or unconsciously, in their choice to use Scots or not. The random forest does not allow us to see the direction this effect takes, but it does suggest that the relationship between *Scots* usage and political affiliation is very robust.

PROFESSION forms the third most important predictor, followed by CREATIVE (whether texts were creative or not). Although both these predictors were particularly important for general literate Scottish society (demonstrated in Figures 10, 11 and 12), they are only the third and fourth most important here. This indicates that different influences were operating on our politically-active authors to those on general Scottish society throughout the eighteenth century. When literate Scottish society was modelled with the random forest and the ctrees, the overall strength of GENRE and PROFESSION was apparent. This can be expected, given such effects would have applied to all authors, rather than a specific subset. The majority of writers in the POLITECS corpus appear to have been influenced first by the genre they were writing in, followed by their profession and thirdly their background and education. Yet Figure 11 has already indicated that the politicians are behaving differently to most of the authors in the corpus, and a closer analysis of this group of writers, along with others harbouring political opinions, has revealed slightly different predictors as most significant.

5.3.4.1.3 Random Forest – Importance Frame

The importance measures presented here have indicated the strength of the numerous predictors influencing the writing of these authors during the eighteenth century, and suggested some interesting and potentially novel factors were in the top positions. However, the variable importance of the different predictors is ranked according to just one kind of importance measure. While Gini decrease is a good measure of variable importance, it can potentially disguise more subtle

relationships between the predictors and the trees. Other importance measures can provide different types of information that can be important to take into consideration. In order to validate whether the ordering of these predictors hold, an importance frame containing various measures of variable importance can also be extracted from the random forest.

For this analysis three types of importance measure were chosen; the *mean minimal depth*, the *Gini decrease* and *times_a_root*. The *mean minimal depth* refers to which level on average the predictor is situated in the ctrees that make up the random forest. The smaller the number, the closer to the root the predictor is situated. *Gini decrease* is same measure used above – the higher the number the greater the node purity. *Times_a_root* indicates how often the predictor is situated at the root of a tree (and thus is voted the most significant predictor). The higher the number, the more often the predictor is in the root. Using these three measures, the importance frame below was extracted from the random forest.

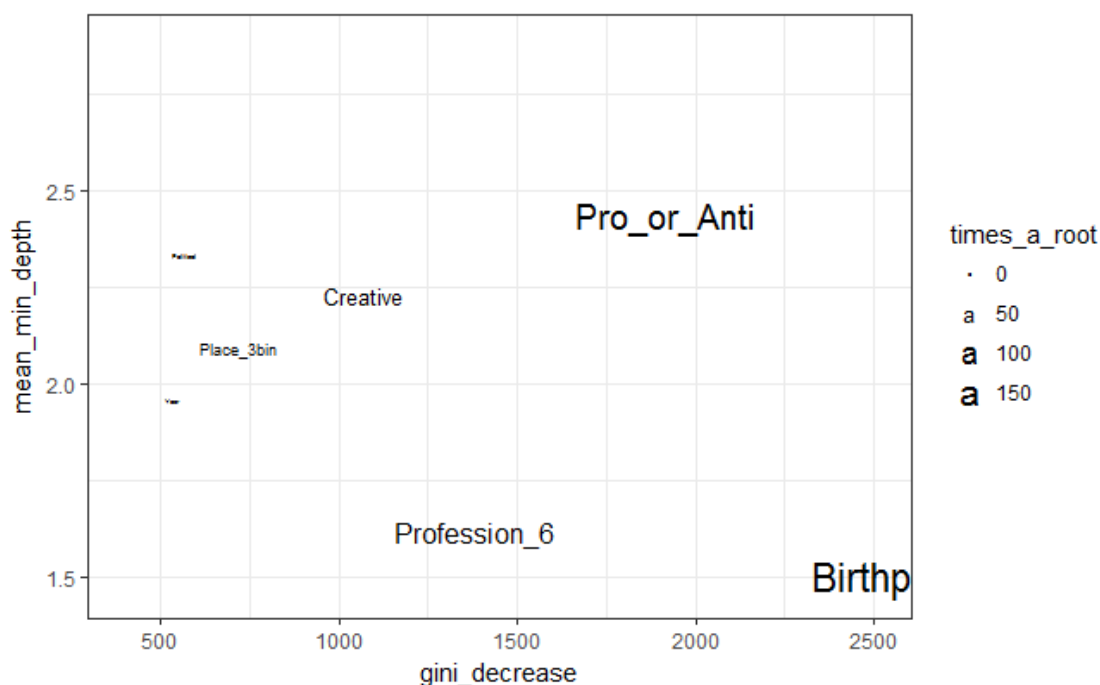


Figure 14: Importance frame showing different factors ranked by three importance measures (mean minimal depth, Gini decreased and times_a_root)

The importance frame confirms that the predictors listed in Figure 13 hold for the data, regardless of which measure of variable importance is selected. The larger the font size of the predictor, and the closer it is to the bottom right corner, the more important it is in determining the variation. It is clear that BIRTHPLACE is still overwhelmingly the most significant factor influencing the language choice of these individuals – it has a high Gini decrease, low minimum depth and high root value. These measures suggest BIRTHPLACE occupies the root of the tree frequently, is often located high up in the tree, and the observations relating to its predictor levels are very pure (either largely *Scots* or largely *English*). It is thus very influential in determining language choice.

Political affiliation (PRO_OR_ANTI) occupies a more interesting position in the frame. It has a high Gini decrease, again suggesting data purity, and a high root value, signalling that it too occupies the root of the tree often. However, it also has a relatively high mean minimal-depth, indicating that on average PRO_OR_ANTI occupies a position further down in the tree. This seems contradictory, but this may be due in part to the high importance of BIRTHPLACE in this dataset. Each of the trees in the random forest contain a subset of the observations, and these observations are themselves a subset of a larger corpus. The effect of a particularly strong predictor can be magnified at such a scale, at the cost of other predictors.

Furthermore, although political affiliation was an important factor influencing these authors, this is not to say that other factors were not influential for particular sub-groups, perhaps especially for those on a particular side of the Unionist divide. This could in turn push PRO_OR_ANTI to occupy a lower position in the ctree, increasing its mean minimal depth overall. The diametric positioning of PRO_OR_ANTI within the importance frame may thus be a reflection the polarity characterising the subset of the data being examined. The situation facing these authors was complicated and perhaps it is not surprising that the predictors follow suit. These importance measures are able indicate the

complex relationship between the predictors and the variable, revealing the multifaceted nature of the multitude of considerations facing politically-active authors writing during this time.

The strength of PROFESSION is also more apparent in the importance frame. Although it does not occur in the root as often as BIRTHPLACE and PRO_OR_ANTI, it is on average positioned high in the tree and has high data purity. Indeed, the importance frame suggests PROFESSION and PRO_OR_ANTI are almost on par in terms of significance. Clearly, PROFESSION played a significant role in determining the choice to use more or less Scots within authors with known political sentiments. Although their BIRTHPLACE and their POLITICAL AFFILIATION were slightly stronger in determining their linguistic decisions, the expectations and restrictions that came with various occupations still played a very important role for these authors. The continuation of Scots in certain fields, including the religious and legal fields could very plausibly be driving some of the effect observed here as well.

Interestingly, YEAR (of publication) and whether the text was political or not are virtually irrelevant for this subset²². Within the eighteenth century, there does not seem to have been any particular clustering of years that encouraged a noticeable increase in Scots, and this is supported by the dendrograms produced in the VNC analysis earlier. Figures 8 and 9 have already indicated the high fluctuations between Scots and English throughout the eighteenth century, but these were likely produced by the idiosyncrasies of individual authors rather than reflecting nationwide trends. The average level of Scots was shown to have been very uniform during this time, both for the general and political populations (though it was elevated for the political population, as demonstrated in Figure 9).

Moreover, POLITECS contains political individuals active at various times throughout the eighteenth century. This includes authors writing during and immediately after the time when the Union agreement took place, but also authors producing work towards the turn of the century, who

²² The importance measures were also extracted from the random forest grown with Year as a raw value, but the difference was almost negligible. Regardless of whether Year was included as a raw value or the VNC clusters, Year as a predictor remained almost irrelevant in conditioning these authors' use of Scots.

reflected upon what had been achieved as a result of the agreement. Whilst there was a rise in political tension towards the second half the century, patriots and political opponents were active throughout. As a result, the lack of an effect for YEAR is perhaps to be expected.

The lack of an effect for POLITICAL (whether the genre of the text was political) is more surprising. These texts do not appear to have especially encouraged the use of Scots for politically-active individuals, despite having considerable influence upon the general literate population. *Political* texts are not the only genre representing these authors; a breakdown of the four recategorized genre levels is represented in Table 4.

Table 4: Number of texts in creative, non-creative, political and non-political categories for subset of politically-active authors

Creative/Not Creative	Political	Not Political
Creative	32	36
Not Creative	41	57

The lack of an effect is thus not being driven solely because *political* texts are the only genre that represent these authors. The politically-active authors come from a variety of backgrounds and professions, and consequently produced a variety of genres. A writer identified as anti-Union in the corpus for example could be represented in writing by a sermon or correspondence, neither of which may have had political overtones and thus would not be coded as *political*. Instead it seems this subset of authors was largely unaffected by the political nature of their texts. It may be that their political affiliation overrides any observed effect relating to the political content of their work, or that the *political* texts examined in general society are largely represented by individuals with political affiliations (which can be expected), who consequently use more Scots lexemes already. It

seems politically-active authors demonstrate higher or lower frequencies of Scots words largely along unionist lines, regardless of the topic.

The various importance measures of the random forest have thus shown that BIRTHPLACE, PRO_OR_ANTI and PROFESSION were the three most important factors predicting the use of Scots words for politically-active authors, confirming that these individuals as they are represented in the corpus did indeed behave differently to the general population represented in POLITECS. Moreover, the strength of political affiliation, as one of these particularly important motivators, suggests that the political changes occurring during the eighteenth century did indeed have an effect on authors' use of Scots. Despite the challenges facing authors writing in Scots during this time, their personal loyalties played a greater role than the restrictions imposed by the genre they were writing in, for example, in their language choices.

The random forest is useful in assessing with greater quantificational accuracy and statistical precision than descriptive statistics alone, just how important a factor such as political affiliation is. However, it is not yet clear in what direction this effect operated. Thus, while the explanations above have been offered to perhaps explain why these factors have been identified as influential, it is as yet unclear whether language use actually follows along these lines. In order to determine whether anti-Union sentiment did encourage more Scots usage, and pro-Union sympathies encouraged more English usage, as we might predict, we need to construct a ctree once more, in order to observe the interactions in the data.

5.3.4.2 Politically-Active Scottish Society – Ctree

Using the subset of politically-active authors, another ctree was grown, using the **partykit** (Hothorn et al., 2010) package in R (R Core Team, 2013). The independent variables included in the model were the same as those used in the random forest, with their recategorized levels (see Figure 13),

except for YEAR, for which raw values were used once again.²³ The dependent variable was Scots words vs English words and AUTHOR and TEXT were included as random effects. The results are shown in Figure 15 below.

²³ This again provided compatibility with the ctrees generated from the general literate audience (Figures 11 and 12). A second ctree using the VNC clusters (as found in Figure 9) instead was also trialed, but this did not boost the positioning of the predictor in the tree, nor did it provide any particular new information to the overall interactions shown in the tree.

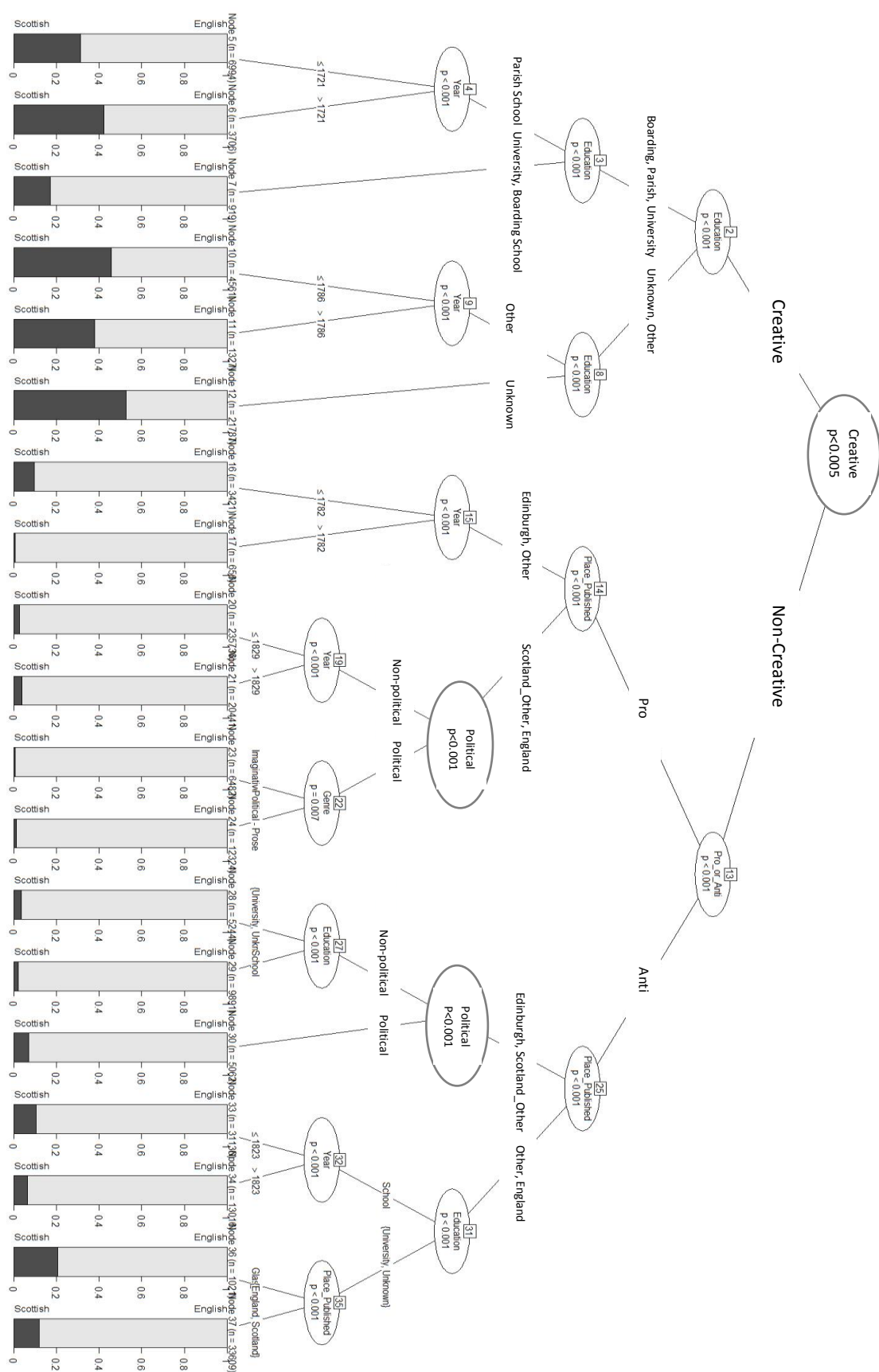


Figure 15: ctree showing all predictors for authors with known political sentiments

It is clear that the top predictors presented in the ctree differ quite substantially from those found by the random forest. This different ordering of importance is the result of the greater quantificational power of a random forest. A random forest is the product of many hundreds of ctrees rather than a single tree. In the case of the random forest grown for the political individuals, this power is aided further by downsampling the data. The ctrees produced here on the other hand is built upon data that is heavily skewed towards *English*, which becomes more problematic when combined with a low number of authors. It is more difficult for the algorithm to identify robust relationships between the importance of the predictors and the minority class with such proportions.

Hence, the ctree indicates that GENRE (*creative* or *non_creative*) is the most important factor conditioning these author's use of Scots, rather than BIRTHPLACE. GENRE no doubt has some importance (it was ranked fourth by the variable importance measures), but it is not the most significant factor, as the ctree suggests. For this reason, among others, we employ a random forest to obtain a more robust assessment of the significance of the relative strength of the predictors. What the ctrees are capable of though, is to identity and display the interactions between these predictors and the variable, even with highly disproportionate data. Accordingly, Figure 15 is able to demonstrate that the divide between higher and lower levels of Scots words by *creative* and *non-creative* writing still applies to this subset of individuals – again we see that proportion of Scots is higher within vernacular literature. It seems political authors did largely follow along the lines set out by the demands of the printing press and the relative freedom to use Scots in vernacular works. The ctree does however, position political affiliation (PRO_OR_ANTI) as a predictor with second-most importance (just as the random forest did). Recall that the closer to the root a predictor is positioned, the higher its importance in determining the variation. A second-level node hence reflects a predictor with high importance. The *non-creative* side is next split by PRO_OR_ANTI, which

then splits into *Pro* (Union) on the left and *Anti* (Union) on the right. Examining the two branches, it is clear that there are higher proportions of Scots within the *Anti* branch than the *Pro* branch.

Although these levels are not as high as those on the *Creative* side of the tree, considering that English had come to dominate the more serious types of prose during the eighteenth centuries, the fact that we are seeing concentrations of Scots at all is an interesting case. Furthermore, these results suggest that those who opposed the Union did in fact use more Scots lexis in their writings than those who supported it. If we compare proportions of Scots in the *Anti* branch to the *Pro* branch, we see almost no Scots present in the latter, with the exception of texts published in *Edinburgh*. The divide between *Pro* and *Anti* is apparent, and the random forest has confirmed that this relationship is robust. The ctree enables us to see the direction that Scots usage follows for political affiliation, and it is clearly along anti-Union lines.

On the *Creative* side of the tree, the second-most important predictor is EDUCATION. Again, this does not match the random forest, for the same reasons given above. The influence of EDUCATION was apparent in the random forest and ctree grown from the general Scottish dataset (see Figures 10 and 11), and it is not unusual to think that the interactions identified between EDUCATION and general society might apply equally well to this subset of authors. Although the proportions of Scots are quite even across the various schooling types, there is notably less Scots usage for those who attended *boarding school* or *university*. This is exactly what we would expect; the English-based, standardising influences of such institutions, their geographical location, the changing expectations regarding the language of academia and the increased contact with English speakers would tend to favour English usage rather than Scots. Creative authors who attended boarding school or university may have had less exposure to Scots and thus were unable to make use of it as a resource for their

works in the same way that many of their contemporaries did, or this may simply reflect personal choice that was guided by their educational past.

This ctree has thus been able to provide the last element to the story of political change and language change in eighteenth century Scotland. It is now clear that political affiliation was very much pertinent to the linguistic choices of the authors with known political sentiments in the corpus. The random forest has indicated the strength of this predictor – it was not just one of many predictors conditioning the variation, but in fact the second most important predictor. The ctree similarly suggests it is a strong predictor, but more importantly has been able to show us the direction of this predictor – and this is in the direction of Anti-Union. Those who were against the Union did in fact use more Scots words, than those who supported it.

5.3.4.3 Results Summary

Taken together, the random forest and the ctree have suggested that the frequency of Scots usage among politically-active individuals was affected by the political and linguistic events of the eighteenth century, that the factors affecting their usage did differ from the general population, and that there is an observable difference between political individuals from either side of the Unionist divide. Specifically; those who were opposed to the Union did use more Scots lexis than those who were in favour of it. The random forest indicated three highly important predictors; BIRTHPLACE, POLITICAL AFFILIATION and PROFESSION. However, the effect of BIRTHPLACE may partially be a product of the sample used. BIRTHPLACE was divided into three simple categories (*Scotland_Other*, *Edinburgh* and *Other*), which could identify marked differences in Scots usage along geographical lines. A number of the politically-active authors in the corpus were born in England (the *Other* category), thus it would be interesting to observe if the effect is still apparent with a larger sample of authors that originated largely from Scotland. This result must thus be interpreted with some caution, and is something to bear in mind for future work. Nonetheless, as the most significant predictor it seems plausible that BIRTHPLACE still had some effect, even if this was partially driven by other factors.

Second to BIRTHPLACE was POLITICAL AFFILIATION, which encouraged the use of Scots along anti-Unionist lines in eighteenth century Scotland. Both the ctree and the random forest measures have suggested the strength of the specific relationship between linguistic choice and the Unionist divide. Finally, PROFESSION has been highlighted as a third key motivator for language choice. Although the positioning of these authors along the political spectrum played a slightly more immediate role in determining their linguistic choices, they were not unaware of the restraints and requirements of their professions. It is surprising not that PROFESSION. made it into the highest-ranked importance measures, but that it was not the top factor. Although it is credible that the turmoil and dissatisfaction that marked eighteenth-century Scotland had an effect on language choice (and this data has indicated that it had a notable effect), we could still expect the restraints of PROFESSION. and GENRE to play the ultimate role. Yet, the results presented here suggest this is not necessarily the case.

6.0 Further Discussion

Clearly there were a multitude of factors operating on Scots usage during the eighteenth century. Depending on whether we are examining a large swathe of Scottish literate society or just a particular sector of it, there were both overlapping and diverging factors that played a crucial role in determining an author's choice to use Scots. As well as the highly specific interplay of factors characterising the position of Scots within literate society, the complexity was multiplied by the nature of the linguistic and political situation at large, with various strands of influence that intersected, intertwined and repelled, creating demands upon the linguistic system that were at times aligned and at times conflicting. This multimodal situation is no doubt the cause for the results presented here; while some effects seem to have been universal across Scottish authors, others seem to have been more specific to particular groups within literate society.

The first part of this research sought to identify how the frequency of Scots lexis patterned over time for the general literate Scottish society, and for politically-active individuals, writing during the eighteenth century (research questions 1 and 2). The dynamic conglomeration of opposing pressures characterising the eighteenth century were already apparent in this temporal exploration of the data, using Variability-based Neighbour Clustering (Gries & Hilpert, 2008), presented in section 5.2. This revealed highly fluctuating data, suggestive of the contrary linguistic and political forces operating in tandem throughout the century to produce movements that contributed both towards linguistic and political conformity to, and divergence from, English and the Union. If we expect that different groups within literate Scotland would be more or less likely to be influenced by various effects, some which they might not share with other groups (and indeed the statistical analysis in this research has suggested that this is the case, at least for politically-active authors), it is not surprising that the temporal analysis has indicated such widely varying levels of Scots and English throughout this time period. Indeed, when the frequency of Scots usage over time was compared

between general society (Figure 8) and politically-active individuals (Figure 9), the VNC models indicated higher average levels of Scots for the latter group. Clearly, different groups of people within Scottish society were reacting to different factors, influences and pressures, as well as the overarching demands made upon all authors throughout the eighteenth century.

The analysis then moved on to address which sociohistorical factors influenced the frequencies of Scots lexis in general literate society, and within the politically-active authors present in the corpus (research questions 3 and 4). The random forest and ctree grown from the data of general literate society (see Figures 10 and 11) suggested that `GENRE` and `PROFESSION` were the most important factors influencing general society. Such broadly applicable effects can be expected to be highly relevant, given that the corpus contains a range of geographically-dispersed text types and authors with widely varying backgrounds, styles and positions within society. With such a diverse range of actors and material, the most influential factor driving any variation in the data would similarly have to be wide in scope and application.

The influence of `GENRE` and `PROFESSION` also reflects the social and historical changes taking place during the eighteenth century. The high prevalence of Scots in creative genres suggests the vernacular backlash that arose during this time, and the renewed acceptability of Scots in creative work as a result of works by influential poets and writers (Clive, 1970; Robinson, 1973; McClure, 1980; Beal, 1997; Jones, 1997; Corbett et al., 2003; Corbett, 2013; Dossena, 2005). The creative literary field was becoming an acceptable channel to use Scots without the suppressive constraints facing authors trying to publish an academic text to a general audience (Dossena, 2005). Levels of Scots were shown to be particularly high in works that were both creative and political (such as *satires* and *radical poetry*), suggesting that the influence of politics did stretch to the cultural sphere. Vernacular literature has been portrayed in various accounts as a literary backlash to the anglicising efforts of the orthoepists (Murison, 1979; Aitken, 1979; Smith, 1970), and poetry was sometimes used as a covert medium by politically-motivated authors (Dossena, 2005; Smith, 2007). It seems

almost consequent therefore, that creative literature and works with political and nationalistic overtones culminate in higher levels of Scots words, relative to creative literature alone.

Non-creative prose on the other hand was influenced by the ongoing anglicisation and the expectations and pressures stemming from publishing interests involving widespread dissemination and the profitability of written works. Such factors encouraged greater use of English in the literature, which makes the noticeable levels of Scots lexis was exhibited by the politicians in the corpus all the more surprising. Despite producing works that were not creative, nor working in a field that was restricted purely to a local level, the proportion of Scots in their writings was considerably greater relative to most other professions represented in the corpus (with the exception of religious and legal professionals as demonstrated in Figure 12).

Their use of Scots lexis, as well as those known to have been for or against the Union, was then examined in more detail to explore whether there was an observable difference between political individuals from either side of the Unionist divide. Specifically, the potential relationship between anti-Union authors and the frequencies of Scots lexemes in their writings, in comparison to pro-Union authors, was of particular interest. This formed the final component of this analysis (research question 5) and again, random forests and ctrees were utilised. By examining the politically-active members of the corpus, the results indicated a statistically significant relationship between Scots usage and Unionist stance. It appears that the turmoil, tension and growing unrest that characterised political change in eighteenth-century Scotland did have an effect on language use, at least for those that engaged with it. Scots lexical items were used more frequently by those that were against the Union, indicating that its use was linked to an anti-Union, nationalistic and patriotic agenda. Resentment or dissatisfaction with the established political order (Phillipson & Mitchison, 1970; Whatley, 2000), a sense or a longing for national pride (MacDonald, 2011), the romanticising of an imagined heroic, independent past (Millar, 2004; Gibbs, 2006), and the rise in patriotism in the wake of the Union (Jones, 1997a; Dossena, 2005), were realised, among other channels, through the

medium of language. There was a growing awareness of Scots and its function as a marker of Scottish identity, brought all the more to light by the zealous efforts of the orthoepists, and this political and linguistic awareness appears to have translated to language use.

Of course, linguistic choices are rarely so tightly defined across distinguishing lines, political or otherwise. Patriotism may not have been wholly tied to an anti-English agenda, and could be realised on both sides of the political spectrum. There were of course those that sought independence and resented English control, but also figures who saw the Union as an opportunity for Scotland's fortunes to improve. Pro-Union did not always necessarily mean pro-English, those that supported the Union did so for many reasons, both pragmatic and strategic (Pentland, 2004). Patriotism can come in many guises, and the clear national boundaries and divisions identified today cannot necessarily be imposed upon historical figures, who may have interacted with political change on diametrical levels. This does not mean that the overall tendency of anti-Union authors to use greater levels of Scots words in their writings is not valid – the statistical models have shown that this relationship is particularly robust. Rather, the complexity of the situation is something to bear in mind when examining this time period.

What is apparent though from these results is an identifiable link between language change and political change in the use of written Scots in eighteenth century Scotland. It seems that those who were involved with a particular political agenda did use language differently to the general society. This suggests that perhaps political change and people's sense of identity and nationalism can be important factors influencing their linguistic choices. This is particularly pertinent to instances of historical language change, given that such analyses are limited to examining only the literate classes of the time, and the literate classes usually included the people actively participating in political life or at least aware of it. As a result, they were liable to form a political opinion or stance, and accordingly be influenced by it. It seems feasible that a similar interaction could be observed in more cases of historical language change, and particular historical instances of profound political change

or unrest could provide new insight into linguistic developments if examined through the lens of patriotism, political resistance and national identity.

Such a relationship is of course plausible in many linguistic scenarios, including contemporary settings. Shoemark et al. (2017) undertook a large-scale sociolinguistic study into the language use of Scottish Twitter users in response to the 2014 Scottish Independence Referendum. They found that users who marked their tweets with pro-independence hashtags tended to use more Scots than those who marked their tweets with anti-independence hashtags. These findings are highly interesting as they seem to align with the results found in this study, suggesting that a similar relationship between the Scots language and political independence operates in contemporary Scotland, more than two hundred years later. However, Shoemark et al. (2017) also found that in general, tweets relating to the independence referendum tended to reflect less Scots than general Twitter activity by the same users. They attribute this difference to style-shifting relative to audience, suggesting that local variants are suppressed when users are trying to reach a broad audience.

Again, similar constraints appear to be operating in contemporary linguistic settings to those present in eighteenth century Scotland. The pressure of reaching a wide audience was a major concern for authors writing in 1700, especially considering the cost of printing and dissemination was far higher than a simple Tweet today. We see the pressure of audience in the results – for the general population *GENRE* formed the strongest predictor, and *non-creative* texts showed very low proportions of Scots, indicating the need to anglicise in order to reach a larger, English speaking readership. Politically-active individuals similarly had lower levels of Scots within *non-creative* genres; audience clearly did play some role in their linguistic choices. Yet the results have also shown that ultimately political affiliation played a greater role in their use of Scots, and Scots is still

observable in noticeable quantities within *non-creative* genres for those who were against the Union.

Concerns with reaching a wider audience and the prevalence of English in Scotland may have become more codified over time, which is what we would expect. Nonetheless, the relationship between politics, language and identity, whilst operating on a different literary platform, is still present in Scotland and does not seem likely to disappear any time soon. What these results have shown, if anything, is that a far-reaching, large-scale socio-political change can in turn filter down affect the most basic and essential component of a coherent social group; their language.

7.0 Limitations and Future Directions

This analysis was naturally limited by the time constraints of the Masters programme, especially given the arduous process involved in creating a historical corpus, based upon texts that require digitisation and transliteration before they can even begin to be used. As a consequence, the final results are the product of a relatively brief foray into this new corpus and there is still the potential for much more research to be undertaken. Furthermore, there were a number of limitations that could not be adequately addressed in the time allocated, and these remain to be resolved at a later date.

The biggest issue was the very low levels of Scots in proportion to English, which makes any kind of quantitative analysis difficult, and resulted in the need to downsample the dataset in order to obtain more accurate results. In future the lists of Scots and English words used to tag words in the corpus could benefit from renewed scrutiny to identify any forms which could be added or removed. For example, function words, which were removed in this analysis, could be included in future research as they may reflect subconscious language choices by the authors rather than deliberate use of features that were salient in Scots or English. It might also be interesting to examine which particular words were favoured by different groups or individuals, for example in the political propaganda issued by the radicals and loyalists.

Secondly, despite spending several weeks and trialling multiple subsets and optimizers on the data, I was unable to get a logistic mixed-effects regression model (Baayen et al., 2008) to converge, without hugely overfitting the model to the data. Unfortunately, generalised linear or logistic mixed effects models can be severely destabilised by inconsistent, imbalanced data, empty cells, collinearity between predictors and highly disproportionate levels of variants (Tagliamonte & Baayen, 2012). There were also periods of time in my data where there was no variation at all, and the majority category (*English*) almost always dominated the minority category (*Scots*), further reducing the ability of the model to find meaningful relationships between the predictors and the

response. Removing periods of zero variation and the reclassification of predictor levels was also trialled, without success. This is the strength of random forests in that they are able to handle these issues, but as a result they are unable to show us the complexities of the interactions in the data in the same way that mixed effects models can. It may be that downsampling is also necessary in order to run more complicated mixed-effects models on the data, but there was simply not the time for this during this project. This is again something to consider for future work.

This analysis also examined only the raw frequencies of Scots lexemes, ignoring features such as specific syntactic constructions or semantic differences between Scots and English. This may not be the best method for examining 'Scottishness' overall, and may have missed considerable swathes of Scots, which could add valuable weight to the proportion of Scots in the dataset. Again, this is difficult to achieve in a corpus study, and previous analyses (Cruickshank, 2012, 2017; Corbett, 2013) examining semantic differences have tended to take a discourse-analysis approach and focussed on a small number of texts. Again, this is something that could warrant further investigation.

A closer look at the political individuals within the corpus could also be very insightful. POLITECS is currently limited in the number of political texts and authors it contains, but this is certainly something to build upon in the future. Of course, this in turn relies upon easier access to historical documents and their digitisation. With a greater time-availability, learning to use specialised historical transcription software such as Transkribus (TRANSKRIBUS team, 2016), could become a feasible option, in turn enabling the growth of the corpus. With a greater number of politically-active authors included it will be interesting to see whether political stance can place them along a linguistic continuum or a linguistic divide. This corpus already holds the potential to facilitate a large amount of qualitative analysis of the data in order to uncover more fine-grained distinctions across language and individuals. With more data and further statistical modelling we will hopefully be able to analyse relationships on multiple linguistic and sociohistorical levels, adding to the picture of Scots development during the eighteenth century.

8.0 Conclusion

Given the time investment required in sourcing, collating and converting texts in order to build a specially-designed corpus, the research undertaken here has been little more than a pilot study into the potentially rich and linguistically-diverse resource now available. Nonetheless, by employing the statistical toolkit and utilising a data-driven approach, this analysis has still managed to provide empirical, informed and robust results that provide fresh insight into the complex dualism characterising eighteenth century Scotland. With newer techniques able to control for the inconsistent information and small sample sizes that so often impedes historical research, we remove the need to artificially impose a focus or presuppose the importance of an effect upon the material we are working with. Instead, we are able to tap into the complex profile of a variable and its many manifestations across different subsets of data. In short, the data tells its own story. And the story uncovered here is one of language, identity and political change in eighteenth century Scotland.

The temporal analysis reflected the antithetical pressures and linguistic attitudes interacting with Scots throughout the eighteenth and into the nineteenth century; levels of Scots and English rapidly fluctuated, creating a plateau in the general decline of Scots during this time. The sociolinguistic analysis utilising non-parametric, decision tree algorithms was able to identify an interaction between language and political change, even with a relatively small sample of authors. Though the vernacular backlash, rise in antiquarianism and the presence of a number of ardent Scottish patriots who disagreed with the Union has been mentioned in various accounts, this linguistic resistance has never been measured. Until now, there had been no quantificational study to determine whether such figures really did use more Scots, and if so, whether their political views were driving this behaviour. This study has been able to show that political affiliation was in fact a very strong factor influencing these individuals' use of Scots, and that linguistic choices were identifiably split along

Union lines, with those who opposed the Union using higher levels of Scots lexemes in their writings, both creative and non-creative.

Of course, most of the political individuals chosen in this sample were recognised as being patriots or outspoken in their political ideas. This safely ensures that they can be labelled and placed on either side of the spectrum. Yet this may mean the anti-Union authors are also more likely to act on this patriotism or nationalistic feeling in their writing, speech and other performances of identity. It is difficult to assess from a corpus study alone whether the use of Scots in these contexts was used deliberately or unconsciously, or a mixture of both, as this requires more detailed, discourse-analysis style research. If their language usage was to some extent deliberate, their choices may be partially personal and partially a confirmation of their public persona. In short, though this still confirms an effect for political affiliation, we must remember that most of these individuals embody the more extreme version of their type. It remains to be seen whether writers who were perhaps less outspoken in their ideas (if these can indeed be sourced) show the same Scots usage along Unionist lines. These subtle differences may perhaps become more obvious with a small-scale, text-by-text analysis, which certainly one of the next steps for this analysis.

Nonetheless, the fact that political affiliation did emerge as strongly as it did in determining their language choices suggests this is a factor that should be taken into consideration in future work examining this time period, as well as perhaps other instances of historic language change that occurred within a backdrop of political change and unrest. Such new and novel results accentuate the comparative value of bagging ensemble algorithms, and the merit of a diverse corpus explored through use of statistical techniques. The combination of robust statistical methods, along with historical insight and a critical understanding of the social and linguistic considerations facing historical actors, can perhaps open new windows on language change, allowing us to further explore the complex playing field of linguistic choice; both public and personal, past and present.

9.0 References

- Adams, J. (1799). *The Pronunciation of the English Language Vindicated from Imputed Anomaly & Caprice: In Two Parts. An Analytical Process Respecting Elementary Combinations and Variations, Chiefly Confined to Monosyllables. An Investigation of Prosody in All the Multiplied Forms of Words, Syllables, Green and Latin Analogy, &c. With an Appendix, on the Dialects of Human Speech in All Countries, And an Analytical Discussion and Vindication of the Dialect of Scotland. By the Rev. James Adams, SRES.*
- Agutter, A., & Cowan, L. N. (1981). Changes in the vocabulary of Lowland Scots dialects. *Scottish Literary Journal*, 14, 49-62.
- Aitken, A.J. (1979). Scottish Speech: a historical view with special reference to the Standard English of Scotland. In A.J. Aitken & T. McArthur (Eds.). *Languages of Scotland* (pp. 85-120). Edinburgh: W&R Chambers.
- Aitken, A. J. (1981). The good old Scots tongue: does Scots have an identity? In E. Haugen, J. D. McClure & D. Thomson (Eds.). *Minority Languages Today* (pp. 72-90). Edinburgh: Edinburgh University Press.
- Aitken, A.J. (1984). Scots and English in Scotland. In P. Trudgill (Ed.) *Language in the British Isles* (pp. 517-532), Cambridge University Press: Cambridge.
- Aitken, A. J. (1990). Address and toast to the Immortal Memory of Robert Burns. In C. Macafee (Ed.). *AJ Aitken: Collected Writings on the Scots Language* (2015, pp. 63-86), [online] Scots Language Centre,
http://media.scotslanguage.com/library/document/aitken/Address_and_toast_to_the_Immortal_Memory_of_Robert_Burns.pdf (accessed 23 Nov 2017).

- Aitken, A.J. (1997). The Pioneers of Anglicised Speech in Scotland; a second look. *Scottish Language*, 16, 1-36.
- Baayen, R. H. (2010). A real experiment is a factorial experiment. *The Mental Lexicon*, 5(1), 149-157.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4), 390-412.
- Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*.
- Bald, M. A. (1926). The anglicisation of Scottish printing. *The Scottish Historical Review*, 23(90), 107-115.
- Bald, M. A. (1927). The pioneers of Anglicised speech in Scotland. *The Scottish Historical Review*, 24(95), 179-193.
- Bald, M. A. (1928). Contemporary references to the Scottish speech of the sixteenth century. *The Scottish Historical Review*, 25(99), 163-179.
- Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.
- Beal, J. (1997). Syntax and Morphology. In C. Jones (Ed.). *The Edinburgh History of the Scots Language* (pp. 335-377). Edinburgh: Edinburgh University Press.
- Bickerton, D. (1973). The nature of a creole continuum. *Language*, 640-669.
- Bickerton, D. (1979). Beginnings. *The genesis of language*, 1-22.
- Blanke, T., Bryant, M., Frankl, M., Kristel, C., Speck, R., Daelen, V. V., & Horik, R. V. (2017). The European holocaust research infrastructure portal. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(1), 1.

- Bono, Paola. (1989). Scottish Studies: Radicals and Reformers in Late Eighteenth Century Scotland. In Horst W. Drescher (Ed.). *Radicals and Reformers in Late Eighteenth Century Scotland: an annotated checklist of books, pamphlets, and documents printed in Scotland 1775-1800*. Frankfurt am Main: Verlag Peter Lang.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Breiman, L., & Cutler, A. (2018). Random Forests (2011). URL http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, Version, 5.
- Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., & Shafait, F. (2013, August). High-performance OCR for printed English and Fraktur using LSTM networks. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on IEEE*, 683-687.
- Brown, P. L., & Levinson, S. S. (1987). *Politeness: Some Universals in Language Use*. Cambridge: Cambridge University Press.
- Brownlee, J. (2015). *8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*. Retrieved from <https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/> (accessed 25 September, 2018).
- Buffoni, F. (1992). *Ramsay e Fergusson precursori di Burns: poesia pastorale e poesia vernacolare nel Settecento scozzese*. Milan: Guerini e Associati.
- Bugaj, J. (2004a). Middle Scots as an emerging standard and why it did not make it. *Scottish Language* 23, 19-34.
- Bugaj, J. (2005). Middle Scots Burgh Court Records: the influence of the text type on its linguistic features. In H. Ritt & H. Schendl (Eds.). *Rethinking Middle English: Linguistic and literary approaches* (pp. 75-88). Frankfurt am Main: Peter Lang.

- Bugaj, J. (2013). The legal language of Scottish burghs. Standardization and lexical bundles (13801560). In R. W. Shuy (Ed.) *Oxford Studies in Language and Law 1* (pp. 337-358). Oxford / New York: Oxford University Press.
- Bukhari, S. S., Kadi, A., Jouneh, M. A., Mir, F. M., & Dengel, A. (2017, November). anyOCR: An Open-Source OCR System for Historical Archives. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on IEEE*, 1, 305-310.
- Burnage, G. (1990). *{\sc celex}: A guide for users*. Nijmegen: CELEX Centre for Lexical Information.
- Caie, G. D. (2007). The Scots language then and now. In J. Sevaldsen & J. Rasmussen (Eds.). *The State of the Union: Scotland, 1707-2007*, Vol. 7 (pp. 21-34). Copenhagen: Museum Tusculanum Press.
- Callender, J. T. (1795). *The Political Progress of Britain*. New York: Richard Folwell.
- Cannon, J. (2015). Fletcher, Andrew, 1655–1716. In J. Cannon & R. A. Crowcroft (Eds.). *A Dictionary of British History*, 3. Oxford: Oxford University Press.
- Clive, J. (1970). The Social Background of the Scottish Renaissance. In N. T. Phillipson & R. Mitchison (Eds.). *Scotland in the Age of Improvement*, 2 (pp. 225-244). Edinburgh: Edinburgh University Press.
- Corbett, J., McClure, J. D. & Stuart-Smith, J. (2003). A Brief History of Scots. In (Eds.). *The Edinburgh Companion to Scots* (pp. 1-16). Edinburgh: Edinburgh University Press.
- Corbett, J. (2013). The Spelling Practices of Allan Ramsay and Robert Burns. In W. Anderson (Ed.). *Language in Scotland: Corpus Based Studies*, Vol. 19 (pp. 65-90). Amsterdam: Rodopi.
- Corpus of Modern Scottish Writing, 1700-1945*. Compiled by Corbett, John & Smith, Jeremy, University of Glasgow, URL: <https://www.scottishcorpus.ac.uk/cmsw/>, (accessed 15 September, 2017).
- Craig, D. (1961). *Scottish Literature and the Scottish People: 1680-1830*. London: Greenwood Press.
- Crawford, R. (1992). *Devolving English Literature*. Oxford: Clarendon.

Crowley, T. (1991) *Proper English? Readings in Language. History and Cultural Identity*, London: Routledge.

Cruickshank, J. (2013). The role of communities of practice in the emergence of Scottish Standard English. In J. Kopaczyk & A. H. Jucker (Eds.). *Communities of Practice in the History of English* (pp. 19-45). John Benjamins: Amsterdam.

Cruickshank, Janet. 2017. 'The Language of Lord Fife in Letters to Lord Grenville 1763-1769'. In Cruickshank, Janet and Robert McColl Millar (Eds.). *Before the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster triennial meeting, Ayr 2015* (pp. 212-231). Aberdeen: Forum for Research on the Languages of Scotland and Ireland.

Daiches, D. (1971). *Sir Walter Scott and his world*. London: Thames and Hudson.

Davidson, Neil. (2003). *Discovering the Scottish Revolution 1692-1746*. London: Pluto Press.

Devitt, A. (1989a). *Standardising Written English: Diffusion in the Case of Scotland, 1520-1659*. Cambridge University Press: Cambridge.

Dictionary of the Scots Language/Dictionair o' the Scots Leid, University of Glasgow. URL: <http://www.dsl.ac.uk/> (accessed 30 March, 2018).

Doetsch, P., Kozielski, M., & Ney, H. (2014, September). Fast and robust training of recurrent neural networks for offline handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on IEEE*, 279-284.

Donaldson, W. (1989). *The language of the people: Scots prose from the Victorian revival*. Aberdeen: Aberdeen University Press.

Dossena, M. (1997). Attitudes to Scots in Burns's correspondence. *Linguistica e Filologia*, 4, 91-103.

Dossena, M. (2000). Sense, shortness and salt: Ideas of improvement in eighteenth-and nineteenth-century collections of Scottish proverbs. *Review of Scottish Culture*, 12, 93-106.

Dossena, M. (2002). A strong Scots accent of the mind: the pragmatic value of code-switching between English and Scots in private correspondence: a historical overview. *Linguistica e filologia*, 14, 103-127

Dossena, M. (2005). *Scotticisms in Grammar and Vocabulary. "Like Runes upon a Standin'Stane?"*. Edinburgh: John Donald.

Dossena, M. (2011). Scottishness and the Book Trade. In S.W. Brown & W. McDougall (Eds.). *The Edinburgh History of the Book in Scotland, Vol. 2: Enlightenment and Expansion 1707-1800* (pp. 1002-1011). Edinburgh: Edinburgh University Press.

Dossena, M. (2012). 'A Highly Poetical Language'? Scots, Burns, Patriotism and Evaluative Language in 19th-century Literary Reviews and Articles. *The Languages of Nation: Attitudes and Norms*, Vol. 148 (pp. 99-118). Bristol: Multilingual Matters.

Dossena, M. (2013). Stour or Dour or Clour: An Overview of Scots Usage in Stevenson's Works and Correspondence. In J. Kirk & I. Macleod (Eds.) *Scots: Studies in its Literature and Language*, 21 (pp. 87-101). Amsterdam: Rodopi.

Douglas, S. (2001). Scots Language and the Song Tradition. In J. M. Kirk & D. P. Ó Baoill (Eds.). *Language Links: The Languages of Scotland and Ireland*, (pp. 233- 236). Belfast: Cló Ollscoil na Banríona.

Fergusson, R. (1773). The GHAISTS: A Kirk-yard Eclogue. *The Edinburgh weekly magazine*, 20, 275-276.

Fischer, A. (2012). *Handwriting recognition in historical documents* (Doctoral dissertation, Verlag nicht ermittelbar).

Fletcher, A. (1706). *State of the controversy betwixt united and separate parliaments*. URL: <https://archive.org/details/stateofcontrover00fletuoft/page/n3> (accessed 30 November, 2017).

- Frank, T. (1994). Language Standardisation in eighteenth-century Scotland. In D. Stein & I. Tiekens-Boon van Ostade (Eds.). *Towards a Standard English, 1600-1800* (pp. 51-62). Berlin: Mouton de Gruyter.
- Fromont, R., & Hay, J. (2008). ONZE Miner: the development of a browser-based research tool. *Corpora*, 3(2), 173-193.
- Fry, M. (1992). *The Dundas Despotism*. Edinburgh: Edinburgh University Press.
- Furber, H. (1931). *Henry Dundas, First Viscount Melville, 1742-1811*. Oxford: Oxford University Press.
- Gibbs, C. (2006). *The New Britons: Scottish Identity in the 18th and 19th Centuries*. Retrieved from http://www.napoleon-series.org/research/society/c_scottishidentity.html#_ednref3
- Görlach, M. (1996). And is it English? *English World-Wide*, 17(2), 153-174.
- Görlach, M. (1998). Even more Englishes: Studies 1996-1997. In E. W. Schneider (Ed.). *Varieties of English Around the World: Text types and the History of Scots Series*, 22 (pp. 55-77). Amsterdam: John Benjamins.
- Gries, S. T. H. (2016d.) *Quantitative corpus linguistics with R. 2nd rev. & ext. ed.* London & New York: Routledge, Taylor & Francis Group.
- Gries, S. T. H. & Hilpert, M. (2008f). The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora* 3(1), 59-81.
- Gries, S. T.H. & Hilpert, M. (2010). Modelling Diachronic Change in the third person singular: a multifactorial, verb and author-specific exploratory approach. *English Language and Linguistics*, 14(3), 293-320.
- Gries, S. T. H. & Hilpert, M. (2012o). Variability-based Neighbour Clustering: a bottom-up approach to periodization in historical linguistics. In T. Nevalainen & E. C. Traugott (Eds.). *The Oxford Handbook on the History of English*, (pp. 134-144). Oxford: Oxford University Press.

- Graham, H. G. (1908). *Scottish men of letters in the eighteenth century*. London: Adam and Charles Black.
- Gordin, M. D. (2015). *Scientific Babel: How science was done before and after global English*. Chicago: University of Chicago Press.
- Hall-Lew, L., Starr, R. & Coppock, E. (2010). Indexing Political Persuasion: Variation in Iraq Vowels. *American Speech* 85(1), 91-102.
- Hall-Lew, L., Frisknew, R. & Scobbie, J. M. (2017). Accommodation or political identity: Scottish members of the UK Parliament. *Language Variation and Change*, 2-40.
- Harris, B. (2005a). Scotland's newspapers, the French revolution and domestic radicalism (c. 1789–1794). *Scottish Historical Review*, 84(1), 38-62.
- Harris, B. (2005b). *Scotland in the Age of the French Revolution*. Edinburgh: John Donald.
- Hay J. B., Pierrehumbert JB., Walker AJ. and LaShell P. (2015) Tracking word frequency effects through 130 years of sound change. *Cognition*, 139, 83-91.
- Hilpert, M., & Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. In M. Kytö & P. Pahta (Eds.). *The Cambridge Handbook of English Historical Linguistics* (pp. 36-53). Cambridge: Cambridge University Press.
- Honeyman, V. (2008). 'A Very Dangerous Place'?: Radicalism in Perth in the 1790s. *Scottish Historical Review*, 87(2), 278-305.
- Hothorn, T., Lausen, B., Benner, A., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in medicine*, 23(1), 77-91.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3), 651-674.

- Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2010). *Party: A laboratory for recursive partitioning*. URL: <http://ftp.auckland.ac.nz/software/CRAN/doc/vignettes/party/party.pdf> (accessed 15 August, 2018).
- Hutchison, G. D. (2017). 'The Manager in Distress': Reaction to the Impeachment of Henry Dundas, 1805–7. *Parliamentary History*, 36(2), 198-217.
- Jack, R. D. S. (1997). The Language of Literary Materials: Origins to 1700. In C. Jones (Ed.), *The Edinburgh History of the Scots Language* (pp. 213-263). Edinburgh: Edinburgh University Press.
- Jander, M. (2016). Handwritten Text Recognition–Transkribus: A User Report. *The electronic Text Reuse Acquisition Project (eTRAP)*.
- Johnson, D. E. (2009). Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass*, 3(1), 359-383.
- Johnston, Paul. (1997). Regional Variation. In Charles Jones (Ed.), *The Edinburgh History of the Scots Language* (pp. 378-432). Edinburgh: Edinburgh University Press.
- Jones, C. (1995). *A Language Suppressed. The Pronunciation of the Scots Language in the 18th Century*. Edinburgh: John Donald.
- Jones, C. (1997). Introduction. In C. Jones (Ed.), *The Edinburgh History of the Scots Language* (pp. 15). Edinburgh: Edinburgh University Press.
- Joseph, John E. (2006). *Language and Politics*. Edinburgh: Edinburgh University Press.
- Kay, P. (1978). Variable rules, community grammar, and linguistic change. *Linguistic variation: models and methods*, 71-83.
- Kay, P., & McDaniel, C. K. (1979). On the logic of variable rules. *Language in society*, 8(2-3), 151-187.
- Kirkham, S. & Moore, E. (2016). Constructing social meaning in political discourse: Phonetic variation and verb processes in Ed Miliband's speeches. *Language in Society* 45, 87-111.

- Kniezsa, V. (1997). The Origins of Scots Orthography. In C. Jones (Ed.), *The Edinburgh History of the Scots Language* (pp. 335-377). Edinburgh: Edinburgh University Press.
- Kopaczyk, J. (2012). Communication Gaps in seventeenth century Britain: Explaining Legal Scots to English Practitioners. In B. Kryk-Kastovsky (Ed.), *Intercultural Miscommunication Past and Present, Warsaw Studies in English Language and Literature* (pp. 217-243). Peter Lang: Berlin.
- Kopaczyk, J. (2013). How a community of practice creates a text community: Middle Scots legal and administrative discourse. In J. Kopaczyk & A. H. Jucker (Eds.), *Communities of Practice in the History of English* (pp. 225-247). John Benjamins: Amsterdam.
- Kuhn, M. (2008). Caret package. *Journal of statistical software*, 28(5), 1-26.
- Labov, W. (1994). *Principles of Linguistic Change, Vol. 1: Internal Factors*. Malden MA: Blackwell.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.
- Llamas, C., Watt, D., & Johnson, D. E. (2009). Linguistic accommodation and the salience of national identity markers in a border town. *Journal of Language and Social Psychology*, 28(4), 381-407.
- Lockhart, G. (1824). *Memoirs Concerning the Affairs of Scotland from Queen Anne's Accession to the Throne: To the Commencement of the Union of the Two Kingdoms of Scotland and England, in May, 1707*. London: J. Baker
- MacDonald, A. A. (2011). The Revival of Scotland's Older Literature. In S.W. Brown & W. McDougall (Eds.). *The Edinburgh History of the Book in Scotland, Vol. 2: Enlightenment and Expansion 1707-1800* (pp. 1012-1025). Edinburgh: Edinburgh University Press.
- MacQueen, L. E. C. (1957). *The Last Stages of the Older Literary Language of Scotland: A Study of the Surviving Scottish Elements in Scottish Prose, 1700-1750, Especially of the Records, National and Local* (Doctoral Thesis, University of Edinburgh, Edinburgh, Scotland). Retrieved from <https://www.era.lib.ed.ac.uk/handle/1842/7316>.

MacQueen, L. E. C. (1983). English was to them a Foreign Tongue, *Scottish Language*, 2, 49-51.

Matheson, C. (1933). *The Life of Henry Dundas, First Viscount Melville, 1742-1811*. London: Constable & Co., Ltd.

Mathison, H. (1995). 'Gude black prent': how the Edinburgh book trade dealt with Burns' Poems. *The Bibliothek; a Scottish Journal of Bibliography and Allied Topics*, 20, 70-87.

Mathison, H. (2007). Robert Burns and National Song. *Scotland, Ireland, and the Romantic Aesthetic*, 77-92.

McArthur, T. (1979). The Status of English in and furth of Scotland. In A.J. Aitken & T. McArthur (Eds.). *Languages of Scotland* (pp. 49-67). Edinburgh, W&R Chambers.

McLean, I., & McMillan, A. (2009). *The concise Oxford dictionary of politics*. OUP Oxford.

McClure, J. D. (1980). Developing Scots as a national language. In J. D. McClure (Ed.). *The Scots Language: Planning for Modern Usage* (11-41). Edinburgh: Ramsay Head Press.

McClure, J. D. (1987). 'Lallans' and 'Doric' in North-Eastern Scottish poetry. *English World-Wide*, 8(2), 215-234.

McClure, J. D. (1994). English in Scotland. *The Cambridge history of the English language*, 5, 23-93.

McClure, J. D. (1996). *Scots and its Literature*. Amsterdam: John Benjamins Publishing.

McCrone, D. (2007). State, Society and Nation: The Problem of Scotland. In J. Sevaldsen & J. Rasmussen (Eds.). *The State of the Union: Scotland, 1707-2007, Vol. 7* (pp. 21-34). Copenhagen: Museum Tusculanum Press.

Meier, A. J. (1997). Teaching the universals of politeness. *ELT journal*, 51(1), 21-28.

Meurman-Solin, A. (1989a). The Helsinki Corpus of Older Scots. Reprinted in *Variation and Change in Early Scottish Prose: Studies Based on the Helsinki Corpus of Older Scots*. Helsinki: Suomalainen Tiedekatemia.

Meurman-Solin, A. (1989b). Variation Analysis and Diachronic Studies of Lexical Borrowing. In G. Caie et al. (Eds.). *Proceedings from the Fourth Nordic Conference for English Studies, 1* (pp. 87-98). Copenhagen: Department of English, University of Copenhagen.

Meurman-Solin, A. (1989c). Variation and Variety in Middle Scots reconsidered: a test study of the Helsinki Corpus of Older Scots. Reprinted in *Variation and Change in Early Scottish Prose: Studies Based on the Helsinki Corpus of Older Scots*. Helsinki: Suomalainen Tiedekatemia.

Meurman-Solin, A. (1992). On the morphology of verbs in Middle Scots: present and present perfect indicative. In M. Rissanen, O. Ihalainen, T. Nevalainen & I. Taavitsainen (Eds.). *History of Englishes. New Methods and Interpretations in Historical Linguistics* (pp. 611-23). Berlin: Mouton de Gruyter.

Meurman-Solin, A. (1993a). *Variation and Change in Early Scottish Prose: Studies Based on the Helsinki Corpus of Older Scots*. Helsinki: Suomalainen Tiedekatemia.

Meurman-Solin, A. (1994). On the Evolution of Prose Genres in Older Scots. *Nowele*, 23, 91-138.

Meurman-Solin, A. (1997a). Differentiation and Standardisation in Early Scots. In C. Jones (Ed.), *The Edinburgh History of the Scots Language* (pp. 335-377). Edinburgh: Edinburgh University Press.

Meurman-Solin, A. (2000b). On the conditioning of geographical and social distance in language variation and change in Renaissance Scots. In D. Kastovsky & A. Mettinger (Eds.). *The History of English in a Social Context. A Contribution to Historical Sociolinguistics. (Trends in Linguistics, Studies and Monographs 129)* (pp. 227-255). Berlin: Mouton de Gruyter.

Meurman-Solin, A. (2003a). Corpus-based Study of Older Scots Grammar and Lexis. In J. Corbett, J. D. McClure & J. Stuart-Smith (Eds.). *The Edinburgh Companion to Scots* (pp. 170-196). Edinburgh: Edinburgh University Press.

Millar, R. M (2003). 'Blind attachment to inveterate custom'. *Language Use, Language Attitude and the Rhetoric of Improvement in the First Statistical Account of Scotland. Insights into Late Modern English* (pp. 311-330). Bern: Lang.

Millar, R. M. (2004). Kailyard, conservatism and Scots in the Statistical Accounts of Scotland. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4(252)*, 163-176.

Millar, R. (2005). *Language, Nation and Power: An Introduction*. Basingstoke, Hampshire: Palgrave Macmillan.

Millar, RM. (2010). An historical national identity? the case of Scots. In D Watt & C Llamas (Eds.), *Language and Identities* (pp. 247-256). Edinburgh University Press, Edinburgh

Millar, RM. (2012). *English Historical Sociolinguistics. Edinburgh Textbooks on the English Language - Advanced*, Edinburgh: Edinburgh University Press.

Millar, RM. (2013). "To bring my language near to the language of men"? Dialect and dialect use in the eighteenth and early nineteenth centuries: some observations. In J.M Kirk & I Macleod (Eds.), *Scots: Studies in its Literature and Language* (pp. 73-87). Amsterdam: Rodopi.

Mitchell, L. C. (2012). Language and national identity in 17th-and 18th-century England. In C. Percy & M. C. Davidson. (Eds.). *The languages of nation: Attitudes and norms, Vol. 148* (pp. 123-140). Bristol: Multilingual Matters.

Morton, G. (1999). *Unionist-nationalism: governing urban Scotland, 1830-1860*, 6. East Linton: Tuckwell Press.

Murdoch, S. (2008). *Anglo-Scottish Culture Clash? Scottish Identities and Britishness, c. 1520-1750*.

Retrieved from <http://revel.unice.fr/cynos/index.html?id=6191>, (accessed 21 December, 2017).

Murdoch, S. & Young, J. R. (2007). Union and Identity: Scotland in a Social and Institutional Context. In J. Sevaldsen & J. Rasmussen (Eds.). *The State of the Union: Scotland, 1707-2007*, 7 (pp. 21-34). Copenhagen: Museum Tusculanum Press.

Murison, D. (1979). The Historical Background. In A.J. Aitken & T. McArthur (Eds.). *Languages of Scotland* (pp. 1-13). Edinburgh: W&R Chambers.

National Archives of Scotland (formerly)/ *The National Records of Scotland*. (2018). URL: <https://www.nrscotland.gov.uk/> (accessed 16 October, 2017).

National Library of Scotland/Leabharlaan Nàiseanta na h-Alba. (2018). URL: <https://www.nls.uk/> (accessed 05 October, 2017).

National Records of Scotland. (2018). URL: <https://www.nrscotland.gov.uk/> (accessed 12 November, 2017)

Nevalainen, T. (Ed.). (1996). *Sociolinguistics and language history: Studies based on the Corpus of Early English Correspondence*, 15. Amsterdam: Rodopi.

Nevalainen, T. (1999). Making the best use of 'bad' data: evidence for sociolinguistic variation in Early Modern English. *Neuphilologische Mitteilungen*, 499-533.

Nevalainen, T. (2006). Historical Sociolinguistics and Language Change. In A. van Kemenade & B. Los (Eds.). *The Handbook of the History of English* (pp. 558-582). Oxford: Blackwell Publishing.

Nevalainen, T., & Raumolin-Brunberg, H. (2000). The third-person singular-(e) S and-(e) TH revisited: The morphophonemic hypothesis. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 130, 235-248.

Nevalainen, T., & Raumolin-Brunberg, H. (2003). *Socio-historical linguistics: Language change in Tudor and Stuart England*. London: Pearson Education.

Oxford English Dictionary. Oxford: Oxford University Press. URL: <http://www.oed.com/> (accessed 14 April, 2018).

Pentland, G. (2004). Patriotism, universalism and the Scottish conventions, 1792–1794. *History*, 89(295), 340-360.

Pentland, G. (2008). *Radicalism, Reform and National Identity in Scotland, 1820-1833*, 65. Suffolk: Royal Historical Society.

Pentland, G. (2011). Pamphlet Wars in the 1790s. In S. W. Brown & W. McDougall, S. W. (Eds.). *Edinburgh History of the Book in Scotland, Volume 2: Enlightenment and Expansion 1707-1800* (pp. 737-752). Edinburgh: Edinburgh University Press.

Pentland, G. (2016). Thomas Muir and the Constitution. In G. Carruthers, & D. Martin (Eds.). *Thomas Muir of Huntershill: Essays for the Twenty First Century* (pp. 185-201). London: Humming Earth.

Phillips, J.D. (2012). Mutual Preservation of Standard Language and National Identity in Early Modern Wales. In C. Percy & M. C. Davidson. (Eds.). *The languages of nation: Attitudes and norms*, Vol. 148 (pp. 123-140). Bristol: Multilingual Matters.

Phillipson, N.T. (1970). Scottish Public Opinion and the Union in the Age of the Association. In N.T. Phillipson and R. Mitchison (Eds.). *Scotland in the Age of Improvement* (125-147). Edinburgh: Edinburgh University Press.

Phillipson, N.T. & Mitchison, R. (1970). Introduction. In N.T. Phillipson and R. Mitchison (Eds.). *Scotland in the Age of Improvement: Essays in Scottish History in the Eighteenth Century* (1-4). Edinburgh: Edinburgh University Press.

Plassart, A. (2014). Scottish perspectives on war and patriotism in the 1790s. *The Historical Journal*, 57(1), 107-129.

Pollner, C. (2000). Shibboleths galore: the treatment of Irish and Scottish English in histories of the English Language. In D. Kastovsky & A. Mettinger (Eds.). *The History of English in a Social Context: A Contribution to Historical Sociolinguistics* (pp. 363-376). Berlin: Mouton de Gruyter.

Riach, W. A. D. (1984). Galloway Schools Dialect Survey. *Scottish Language*, 3, 49-59.

Robinson, M. (1973). Modern Literary Scots: Fergusson and after. In. A. J. Aitken (Ed.). *Lowland Scots: Papers Presented to an Edinburgh Conference [held on 12-13th May 1972] by The Association for Scottish Literary Studies Occasional Papers No. 2* (pp. 38-49). Edinburgh: Edinburgh University Press.

Robinson, M. (1983). Language choice in the Reformation: The Scots Confession of 1560. In J. D. McClure (Ed.). *Scotland and the Lowland tongue Studies in the Language and Literature of Lowland Scotland* (pp. 59-78), Aberdeen: Aberdeen University Press.

Romaine, S. (1982). *Socio-Historical Linguistics: its status and methodology*. Cambridge: Cambridge University Press.

Sankoff, D. (1975). VARBRUL 2. *Unpublished program and documentation*.

Scott, P. (1992). *Andrew Fletcher and the Treaty of Union*. Edinburgh: John Donald.

Shoemark, P. Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 1*, 1239-1248.

Simistira, F., Ul-Hassan, A., Papavassiliou, V., Gatos, B., Katsouros, V., & Liwicki, M. (2015, August). Recognition of historical Greek polytonic scripts using LSTM networks. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on IEEE*, 766-770.

Smith, D. (forthcoming). 'The predictability of {S} abbreviation in Older Scots manuscripts' In R. Alcorn, B. Los, J. Kopaczyk and B. Molineaux, (Eds.). *Historical Dialectology in the Digital Age* Edinburgh: Edinburgh University Press.

Smith, J. A. (1970). Some Eighteenth-Century Ideas of Scotland. In N.T. Phillipson and Rosalind Mitchison (Eds.). *Scotland in the Age of Improvement* (pp. 107-124). Edinburgh: Edinburgh University Press.

Smith, J. J. (1996). Ear-rhyme, eye-rhyme and traditional rhyme: English and Scots in Robert Burns's Brigs of Ayr. *Glasgow Review*, 4, 74-85.

Smith, J. J. (2007). Copia verborum: The linguistic choices of Robert Burns. *The Review of English Studies*, 58(233), 73-88.

Szechi, D. (1997). Constructing a Jacobite: the social and intellectual origins of George Lockhart of Carnwath. *The Historical Journal*, 40(4), 977-996.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language variation and change*, 24(2), 135-178.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1), 24-54.

Team, R. C. (2013). *R: A language and environment for statistical computing*. Vienna: The R Foundation for Statistical Computing. URL: <http://www.R-project.org/>

Templeton, J. M., & Aitken, A. J. (1973). Lowland Scots; papers presented to an Edinburgh conference [on 12th-13th May 1972] (No. 2). *Association for Scottish Literary Studies*.

Trudgill, P. (1986). *Dialects in contact*. Oxford: Blackwell.

The Helsinki Corpus of Older Scots (1995). Department of Modern Languages, University of Helsinki. Compiled by Anneli Meurman-Solin.

The People's Voice (2018). Led by MacDonald, C., Blair, K. & Carruthers, G. University of Glasgow & University of Strathclyde, Glasgow. Retrieved from <http://thepeoplesvoice.glasgow.ac.uk/project-team/> (accessed 09 December, 2018).

van Eyndhoven, S. & Clark, L. (forthcoming). The <quh-> - <wh-> switch: an empirical account of the anglicisation of a Scots variant in Scotland during the sixteenth and seventeenth centuries. *English Language and Linguistics*.

Wagenknecht, E. (1991). *Sir Walter Scott*. London & New York: Continuum International Publishing Group.

Watts, R. J., Ide, S. & Ehlich, K. (1992). *Politeness in Language. Studies in its History, Theory and Practice*, (Eds.). Berlin/New York: Mouton de Gruyter.

Whatley, C. A. (2000). *Scottish Society, 1707-1830: beyond Jacobitism, towards industrialisation*. Manchester: Manchester University Press.

Wickham, H. (2015). stringr: Simple, consistent wrappers for common string operations. *R package version, 1(0)*.

Wilkinson, D. (2002). COCKBURN, John (c.1679-1758), of Ormiston, Haddington. In S. Handley, D. Hayton, E. Cruickshanks (Eds.). *The History of Parliament: the House of Commons 1690-1715*. Woodbridge: Boydell and Brewer.

Witten, D., James, G., Tibshirani, R., & Hastie, T. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Wright, M. N., & Ziegler, A. (2015). Ranger: a fast implementation of random forests for high dimensional data in C++ and R, *arXiv preprint arXiv:1508.04409*.

Acknowledgements

First and foremost, this Masters could not have been completed without the tireless effort, enthusiasm, patience, encouragement and support from my supervisor, Dr Lynn Clark. Thanks to her willingness to go along with my sometimes overly ambitious ideas and plans, she allowed me to navigate my way through a wonderful year spent exploring the complexities of eighteenth century Scots, gently guiding me and offering motivation when it was needed so that we could eventually see the hypothesis come to light. I enjoyed our endless chats and brainstorming sessions, and I will certainly miss her constant positivity and valuable insight greatly!

Much gratitude must also be extended to Dr Vica Papp, the computer wizard who helped me find a solution to every technological problem and whose advice and guidance have taught me many new valuable skills (and helped me acquire a new level of patience!). The many hours spent together cleaning the data and collating results were priceless to this investigation, and without her help such impressive results would surely not have come to light. I would also like to thank Robert Fromont for his endless patience and help as we attempted to build a new corpus and search its contents. Robert provided many good suggestions and enabled me to work closely with him to create the end product I desired. His LaBB-CAT skills and insightful implementations to the corpus were invaluable to this research also.

Furthermore I would like to thank Vicky Watson for her endless optimism, enthusiasm, encouragement and editing advice, as well as the many wonderful tea and coffee breaks that kept me going to the end. Also thanks to Andrew Curtis-Black for helping me to automate some of the wordlist collation process, and generally providing all-round support and happiness to help me get through the masters. Many thanks as well to Jennifer Middendorf and Ryan Podlubny for helping me to scan and OCR my historical documents using their special software programmes, to Wendy Anderson for sending through the Master Spreadsheet I required for the Corpus of Modern Scottish

Writing, to Stefan Gries for sending me the VNC script and to Yvonne Shand for processing my rather complicated order with the National Library of Scotland and ensuring the correct pages and information were sent through.

I also wish to thank Dr Heidi Quinn, Dr Rhona Alcorn, Dr Jane Stuart-Smith, Ryan Podlubny, Sidney Wong, my family members, the associated members of the NZILBB, the Syntax Socio reading group, and the University of Canterbury Linguistics Department, for their thoughts, suggestions, improvements and inspiration, from which this thesis has benefitted considerably. Finally, many thanks to Dr Kevin Watson and Dr Joanna Kopaczyk for their time and energy committed to examining this thesis and providing their own insights.

Appendices

Appendix One

Political Author	Documents
George Lockhart	<p>The Lockhart Papers and Memoires</p> <ul style="list-style-type: none"> • <i>Memoirs concerning Scotland, 1707-1708</i> • <i>Memoirs concerning the affairs of Scotland</i>
Henry Dundas	<p>Correspondence of Henry Dundas</p> <ul style="list-style-type: none"> • Henry Dundas to Lord Chancellor, 1793 • Letter of Henry Dundas to advocate (lawyer), 1796 • Correspondence of Henry Dundas 1817 • Correspondence of Henry Dundas, 1771 • Correspondence of Henry Dundas 1781
Sir Walter Scott	<p>The Letters of Sir Walter Scott: E-Text</p> <ul style="list-style-type: none"> • 1787-1807 • 1808-1811 • 1812-1817 • 1818-1825 • 1826-1832
John Cockburn	<ul style="list-style-type: none"> • East Lothian agricultural bibliography. • Representation to his Grace Her Majesties High Commissioner, and the right honourable Estates of Parliament, for John Cockburn younger of Ormistoun.
Alexander Rodgers	<ul style="list-style-type: none"> • The alter of liberty, or, Songs for the people • Clerical anecdotes, and Parson's comic songster: advice to the priest-ridden, also, a joiner's bill • Poems and songs, humorous and satirical.
Andrew Fletcher	<ul style="list-style-type: none"> • An account of a conversation concerning a right regulation of governments • Letter concerning Home rule for Scotland: as advocated by Andrew Fletcher of Saltoun: with its bearing in support of Home Rule for Ireland. • An historical account of the ancient rights and power of the Parliament of Scotland: to which is prefixed, a short introduction upon government in general • The political works of Andrew Fletcher, Esq; of Saltoun.

Appendix Two

Correspondence of Sir Charles Gilmour, bt.	<ul style="list-style-type: none"> • Andrew Fletcher to Charles Gilmour, 1730. • John Cockburne to Sir Robert Walpole, 1732 • Cockburn, addressed 'sir' Sir Charles Gilmour, n.d • John Cockburn to Sir Charles Gilmour, 1740 • John Cockburn to Sir Charles Gilmour, 1741 • John Cockburn to Sir Charles Gilmour, 1742
Miscellaneous correspondence addressed to Sir Charles Gilmour	<ul style="list-style-type: none"> • James Erskine of Grange; William Pulteney; Tweeddale; Lord Graham; John Cockburn; Alexander Cuninghame of Bonnington; John Marjoribanks, Hallyards.
Papers of the Graham Family, Dukes of Montrose (Montrose Muniments)	<ul style="list-style-type: none"> • Correspondence of James, 1st Duke of Montrose: John Cockburn, son of Adam Cockburn of Ormiston. London • Letters to Mungo Graham of Gorthie: John Cockburn, 1712-1715 • Correspondence and personal papers, 1566 – 1941 • Letter to Lord Grange from John Cockburn, yr, of Ormiston, in London, asking for his vote at the next Parliamentary election, 1722

Appendix Three

Predictor	Predictor Levels	Reclassified as
Genre	Verse/Drama Imaginative Prose Political – Creative Orthoepist	Creative
	Administrative Prose Political - Prose Instructional Prose Journalism	Non-Creative

	Personal Writing Expository Prose Instructional Prose Correspondence – Political Religious Prose	
Genre	Political – Creative Correspondence – Political Political - Prose Orthoepist Verse/Drama Imaginative Prose Administrative Prose Instructional Prose Journalism Personal Writing Expository Prose Instructional Prose Religious Prose	Political Non-Political
Profession	Politician Author Poet Legal Professional Orthoepist Other	Politician Author_Creative Author_Non-Creative Poet Religious/Legal Professional Orthoepist Academic Other
Education	Boarding School Parish School University Unknown Apprenticeship Secondary School	Boarding School University Parish School Other
Birthplace	Glasgow Edinburgh Scotland_Other England France Aberdeen Unknown	Scotland_Other Edinburgh Other
Place Published	England Glasgow Unknown Edinburgh Scotland_Other America Australia Europe Ireland	Scotland_Other Edinburgh England Other

List of Figures

Figure 1: Home Page of LaBB-CAT for POLITECS.....	59
Figure 2: The search page of LaBB-CAT containing some of the various annotation layers for POLITECS	60
Figure 3: LIWC manager tagging a text with words identified from the Scots and English dictionaries	63
Figure 4: Participant attributes for example participant (Sir Walter Scott) in LaBB-CAT	66
Figure 5: Various filters applied to a simple orthographic search string in LaBB-CAT which removes all Standard English words.....	75
Figure 6: LaBB-CAT's search page with various filters checked to enable a search for Standard English words only.....	83
Figure 7: Search string using LIWC category.....	86
Figure 8: VNC analysis showing entire corpus with thirteen clusters.....	99
Figure 9: VNC of levels of Scots words for those known to have pro or anti Scots sentiments (5 clusters).....	103
Figure 10: Variable Importance measures from random forest for general literate Scottish society	114
Figure 11: ctree showing proportions of Scots across general literate society with all predictors included.....	120
Figure 12: ctree showing proportions of Scots for general literate Scottish society with genre recoded to include Political texts.....	126
Figure 13: variable importance of factors predicting use of Scots for politically-active authors	133
Figure 14: Importance frame showing different factors ranked by three importance measures (mean minimal depth, Gini decreased and times_a_root).....	136
Figure 15: ctree showing all predictors for authors with known political sentiments	142