English Language and Linguistics, 24.1: 211–236. © Cambridge University Press 2019
doi:10.1017/S1360674319000078

The <quh->–<wh-> switch: an empirical account of the anglicisation of a Scots variant in Scotland during the sixteenth and seventeenth centuries[1]

SARAH VAN EYNDHOVEN and LYNN CLARK
University of Canterbury
(Received 12 March 2018; revised 7 October 2018)

This article explores the anglicisation of the Scots language between the sixteenth and eighteenth centuries, focusing on the variation between the orthographic clusters <quh-> and <wh-> found in relative and interrogative clause markers. Using modern statistical techniques, we provide the most comprehensive empirical analysis of this variation so far in the *Helsinki Corpus of Older Scots* (Meurman-Solin 1995). By combining the techniques of Variability-Based Neighbour Clustering (Gries & Hilpert 2008, 2010, 2012) with mixed-effects logistic regression modelling (Baayen *et al.* 2008), we uncover a different trajectory of change than that which has previously been reported for this feature (Meurman-Solin 1993, 1997). We argue that by using modern methods of data reduction and statistical modelling, we can present a picture of language change in Scots that is more fine-grained than previous studies which use only descriptive statistics.

Keywords: historical Scots, anglicisation, quantitative corpus analysis, statistical modelling

1 Introduction and background

1.1 Historical developments

From its first literary appearance in 1375, until the mid 1500s, Scots was the dominant language variety of Scotland. Some characterise it as a fully functioning language with an emerging standard and a wealth of literature (Romaine 1982; Devitt 1989; Pollner 2000; Douglas 2001); others suggest that English and Scots existed on a linguistic continuum (Aitken 1984a; Görlach 1996; Kniezsa 1997; Kopaczyk 2012), with a large common lexical core (Meurman-Solin 1993). Regardless, the 'heyday' of Scots (Murison 1979: 8-9) was disrupted by various social and political developments in the sixteenth to eighteenth centuries and Scottish Standard English (SSE) developed as the new nationwide standard. This anglicised standard became preferred in the professional arena and most written genres.

[1] This article has benefited considerably from the comments made by two anonymous *ELL* reviewers and by the editor Patrick Honeybone. We are very grateful to them for their time and positive feedback, which has improved this article considerably. We also received useful feedback from the presentation at the 2015 New Zealand Linguistics Society Conference in Dunedin, New Zealand. Furthermore, we would like to thank Vicky Watson and Liam Walsh for their feedback, helpful comments and support. All remaining errors and shortcomings are very much our own.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

The rise of SSE has been linked to a number of sociohistorical events that occurred during the sixteenth to eighteenth centuries. In particular; printing (Devitt 1989; Kniezsa 1997; King 1997; Douglas 2001), religious upheaval (Aitken 1979, 1984a; Bugaj 2004; Millar 2005; Lawson 2014) and the Union of the Crowns (Aitken 1979; Pollner 2000; Douglas 2001) have been identified as keystones in the anglicisation of Scots. Yet it is easy to view events as a historical narrative leading up to some ultimate goal: in this case, the union of two nations (Kopaczyk 2012). The period was by no means harmonious, and Scottish society was far from homogeneous in its feelings towards the Union. It seems unlikely that all sections of society rapidly shunned Scots (Meurman-Solin 1993). Indeed, institutions at a purely local level tended to resist anglicising tendencies, regarding vernacular features as prestigious (Meurman-Solin 1993; Kopaczyk 2012). Despite the suddenness of the political and social changes facing the people of Scotland at this time, the switch from Scots to English is reported to have been more gradual, with overlapping processes of divergence and convergence to/from Scots throughout this time period. For instance, Meurman-Solin (1997) has shown that while some contemporary texts displayed a rapid decrease in Scots variants, others showed an increase. Authors made the switch from Scots to anglicised forms at different times for different Scots features. Indeed, the prevalence of a 'mixed dialect' (McClure 1983; Aitken 1984a) or 'mixed speech' (Meurman-Solin 1997) has been observed throughout seventeeth-century Scots literature.

The decline of written Scots and the subsequent rise of SSE has been extensively documented thanks to a wealth of literary material spanning a sizeable historical time frame. This research has revealed a complex combination of social, political and textual constraints operating on the emergence of SSE.

1.2 Previous studies

The first large-scale study on the anglicisation of Scots was undertaken by Devitt (1989), who looked at five Scots variables and their anglicisation across text type (genre) and time. She suggested that the set conventions and expectations surrounding different text types could explain their levels of anglicisation. Following this, Meurman-Solin (1989a, 1989b, 1989c, 1992, 1993, 1997, 2003) has undertaken by far the most detailed analysis of Scots and SSE. Utilising the extensive *Helsinki Corpus of Older Scots* (HCOS), she has been able to produce a wealth of research that examines with greater breadth the complex historical factors at play than previous small-scale studies. She characterises the rise of SSE as alternating between periods of rapid and slow change (1993), and has revealed considerable heterogeneity in the appearance of features across text, time period and author (Meurman-Solin 1989b, 1989c, 1993). A range of different social and contextual factors have been explored in her research, including TEXT TYPE (1992), AUDIENCE (1992, 1997), style (labelled 'CONTEMPORANEITY') (1989c) and TEXT MEDIUM (*printed* or *manuscript*) (1992, 1997). These factors may act independently or in conjunction, and various conservative or innovative forces were particularly relevant at different times in the move to standardisation (Meurman-Solin 1992, 1993).

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

In most quantitative work on the history of Scots, specific social factors are correlated with particular linguistic features in a piecemeal way. This creates an artificial and arbitrary sense of the separateness of these social constraints on the changes which took place, and has resulted in differing and sometimes conflicting claims concerning the key factors constraining or facilitating the anglicisation of Scots. For example, Devitt (1989) only examined the correlations between genre and time, which may have disguised patterns stemming from other sociohistorical factors. Meurman-Solin (1993) has focused mainly on text type, although she has acknowledged that other socially conditioned factors such as audience could have influenced anglicisation. Furthermore, both these authors base their conclusions on a discussion of descriptive statistics rather than by means of a stepwise multiple regression, which would compute the significance of one independent variable (e.g. genre) while explicitly controlling for the effects of all other known independent variables (e.g. audience, text medium, style).

Romaine (1982) recognised the untapped potential of using regression modelling in the diachronic study of Middle Scots. Using the variable rule program VARBRUL (Sankoff 1975), Romaine found that alongside linguistic environment, stylistic constraints in particular influenced < wh- > relative deletion. Romaine (1982) was able to observe the effects of multiple (rather than individual) predictors on historical data simultaneously, representing a significant step forward in historical Scots research. However, since this publication in 1982, there have been considerable advances in statistical modelling of variable data. For instance, Romaine (1982) was only able to examine one set of extralinguistic constraints at a time (so called 'fixed effects'), because in the original VARBRUL program there was no way to explore interactions between the independent variables (or 'factor groups'). Furthermore, there was no way to account for the seemingly random variation that is always present in a dataset, which could be attributed to the idiolect of particular authors or the trajectory of change of individual words (for more on this and the constraints of VARBRUL, see Johnson 2009). Indeed, Romaine (1982: 207) herself noted that 'the multivariate analysis may conceal as much as it can reveal'. Despite the advances made by Romaine (1982), research on Scots has not moved in the direction of incorporating newer and more sophisticated methods of statistical modelling (although see Smith (forthcoming) for a statistical analysis of spelling variation in fifteenth-century Scots).

The need for approaches utilising not only greater statistical accuracy, but also the capacity to recognise the multifarious and heterogeneous nature of historical data, is becoming clear. Previous studies on Scots have not adopted techniques that adequately come to grips with the huge variability of diachronic corpora, nor the antipodal pressures stemming from local and supraregional interests. Yet this has become more achievable with modern methods of modelling variation, creating greater scope to pinpoint possible factors influencing a particular instance of language change. We argue that by adopting current statistical modelling techniques, we can reach a better explanatory account of the factors which promoted or inhibited language change in Scotland. Accordingly, we adopt some of these newer empirical methods as we re-

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

examine variation in the orthographic clusters <quh-> and <wh-> occurring in relative and interrogative pronouns, in the HCOS.

2 Methods

2.1 Circumscribing the variable

The Scottish orthographical variant <quh-> corresponds to the initial <wh-> cluster in English, in relative and interrogative pronouns such as *which*, *where*, *what*, *whom*, which in Old Scots were represented as *quhilk*, *quhere*, *quhat*, *quhom.* During the seventeeth century <quh-> came to be replaced with the anglicised variant. However, spelling practices were not standardised during the sixteenth and seventeeth centuries and there was substantial variation in this variant, including <qu->, <qw->, <qwh-> and <qh->, though <quh-> was by far the most common. Furthermore, during the switch to <wh->, 'transitional' spellings combining a mixture of Scots and English orthographic variants have been identified (Kniezsa 1997; Beal 1997). However, Laing & Lass (forthcoming) suggest that the different spellings were not purely the result of orthographic variation but corresponded to specific phonological realisations distinguishing Scottish and northern English dialects from southern English. Evidence from various historical corpora suggests the orthography represented a phonological distinction between northern [kw] for <qu-> spellings, [xh] for <quh-> (which later changed to [hw] after historical processes of lenition occurred), and southern [w] for <wh-> (Laing & Lass forthcoming). Thus the change from <quh-> to <wh-> may not be purely orthographic, but also reflective of historical phonological changes taking place, coupled with the influence of English on Scots. In this article we will not explore the phonological implications of this change, basing our examination purely on orthographical variation. However, the distinction is important to keep in mind.

<quh-> has been included in many studies, no doubt due to its emblematic nature as a Scots variant that underwent clear and unambiguous anglicisation. However, Lass & Laing (2016) and Laing & Lass (forthcoming) have identified the various spelling variants of 'qu' occurring in Early and Late Middle English, and Kniezsa (1997) has noted that <quh-> was the usual spelling for the extreme north of England as well.[2] Nonetheless, it seems <quh-> was vastly preferred for Older and Middle Scots, unlike Old and Middle English (Lass & Laing 2016).

Studies that have specifically examined <quh->/<wh-> have noted the categorical nature of the switch. Devitt's (1989) analysis pinpointed the year 1600 as pivotal; use of <quh-> decreased dramatically whilst <wh-> moved from 17 to 83 per cent usage. Most of Devitt's (1989) texts exhibited categorical use of <wh-> or <quh-> and

[2] Kniezsa (1997) has identified the counties of Cumberland, Northumberland, Durham, Lancashire, Westmoreland and North Riding as similarly using this variant, with the exception of York. <qu> as a variant only became regular in English from the thirteeenth century onwards (Blake 1992), and is thought to represent [kw]-initial words of Germanic origin (Lass & Laing 2016). <qu-> had, however, become a minor spelling variant by the fifteenth century.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

diffusion across texts was strongly suggestive of an s-curve pattern of change. Meurman-Solin's (1997) analysis of the HCOS also suggested a rapid decrease in <quh-> and rise in <wh->, though right up until 1700 there is considerable oscillation across texts. It seems that either <quh-> or <wh-> was preferred in a text, rather than any kind of variable usage (Devitt 1989; Meurman-Solin 1997). Though both studies have indicated similar results, Meurman-Solin's findings are of most interest as this study will also seek to use the HCOS to analyse the <quh-> cluster.

The data for this project come from the *Helsinki Corpus of Older Scots* (HCOS; 1995). This corpus of 850, 000 words of running text is the largest computer-readable corpus of Older Scots texts. It contains edited texts or early prints from a wide range of genres including Acts of Parliament, local records, trial proceedings, sermons, pamphlets, scientific and educational treatises, histories, biographies, diaries and private letters. Using AntConc (Anthony 2015, version 3.5.0) to search the HCOS text files, the clusters <qu->, <quh->, <qw->, <qwh-> and <wh-> <vh-> and <hw> were included in the search string. We included all the more common variants of 'qu' and 'wh' to incorporate a wider range of variation, given that orthographic practices were variable at the time (and phonological changes were also taking place, see Lass & Laing 2016). The vast majority of tokens were <quh-> and <wh-> with very few hits for the remaining clusters (38 tokens altogether). These were subsequently relabelled as <quh-> or <wh-> and merged with the respective datasets. The results were then circumscribed; ambiguous or unknown tokens were checked using the online *Dictionary of Scottish Language* (*DSL*; www.dsl.ac.uk) and invalid tokens were removed. Incomplete tokens, often marked wh~ in the corpus, were deleted. Furthermore, <quh-> was used categorically before 1570 so we also removed all tokens occurring between 1450 and 1569. This left 7,759 potential sites of variation to explore.

A number of extralinguistic variables are encoded into each text file in the HCOS, including PUBLISHING DATE, AUDIENCE, CONTEMPORANEITY (the style of the writing), TEXT MEDIUM [*printed*, *manuscript*], LITERARY MEDIUM [*script*, *speech-based*, *written*], the author's RANK, AGE and SEX, INTERACTIVENESS (whether the text was designed to engage the reader or simply state facts), and RELATIONSHIP TO THE ADDRESSEE [*intimate, distant*] (for letters). However, the amount of available information varies widely for different texts and a degree of manual annotation was often necessary. For example, in the case of court proceedings, the texts were carefully read to try to determine the author of each token produced. If the variant came from a speaker who was being directly quoted, they were marked as author, but if the variant came from a scribe who was narrating the series of events, the author was listed as 'unknown'.

The HCOS has been divided into four time periods: 1450-1500, 1500-70, 1570-1640 and 1640-1700. Meurman-Solin (1989a) acknowledges that the time periods of the corpus do not correspond to key diachronic developments in the history of Scots; rather they have been chosen to match the time periods of the *Helsinki Corpus of English Texts* (Rissanen *et al.* 1991). While this may be convenient for comparing developments in the history of English and the history of Scots, it is not driven by how the data themselves pattern over time. When we are examining the variation and

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

change of a single linguistic variable, we should perhaps be mindful that the variable itself may have its own textual history. To avoid distorting or disguising that, we need a better way to model variation over time. Accordingly, this study employs the technique of Variability-Based Neighbour Clusters (Gries & Hilpert 2008) to explore change over time in <quh-> and <wh->, which we explain in the course of the next section.

3 Analysis

3.1 VNC

Historical analyses of language variation and change tend to describe trajectories of change using pre-set, equal-length time periods that are artificially imposed on the data, as a result of subjective categorisation made by corpus compilers (as is the case in the HCOS) or based on well-established time periods that have been defined by key sociohistorical changes (Gries & Hilpert 2010). Yet sectioning the data into convenient year-frames can disguise or overlook trends, painting an incomplete picture of the subtle changes that may characterise the trajectory of any one variant. Trends, turning points and slopes can all be altered or missed when such categorisation is applied (Gries & Hilpert 2010). Traditional period divisions can also mask non-linear developments, and this periodization can discourage research across these convenient boundaries (Nevalainen 2006). Furthermore, sectioning the data according to major historical events ignores the time lag that may ripple through language change.

Gries & Hilpert (2010) thus developed a statistical method to section temporal data: Variability-Based Neighbour Clusters (VNC). This can be used to determine coherent temporal stages as well as conservatively identify data points as outliers. In the model, data are fed into the algorithm which then determines which data points cluster most closely together. Clusters are defined by a high level of within-group similarity and low level of across-group similarity. The measurement of similarly can be set to generate clusters that constitute a relatively homogenous period of interest. The data, rather than the researcher, determine the temporal stages, hence the periods are derived directly from the phenomenon under investigation (Gries & Hilpert 2010). This is a step towards a more accurate, quantitatively constructed analysis of historical data by removing the need for arbitrary divisions, such as those imposed in the HCOS.

The first stage of this research was to run a VNC analysis on the *quh ~ wh* variable data to explore how the frequency of <wh-> clustered over time. The VNC algorithm is available as an R (version 3.1.2, R Core Team, 2012) script, which was kindly sent to us by Stephan Gries (p.c). The results of this analysis are presented in the dendrogram in figure 1.

The y-axis indicates the difference in standard deviations from the mean frequencies of <wh-> in each of the merged temporal files. The x-axis indicates chronological year from 1570 to 1707. Hierarchical clustering algorithms typically cluster similar data together; the difference here is that the clustering algorithm also pays attention to the time depth of the data so that the clusters are grouped not only by similarity in

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

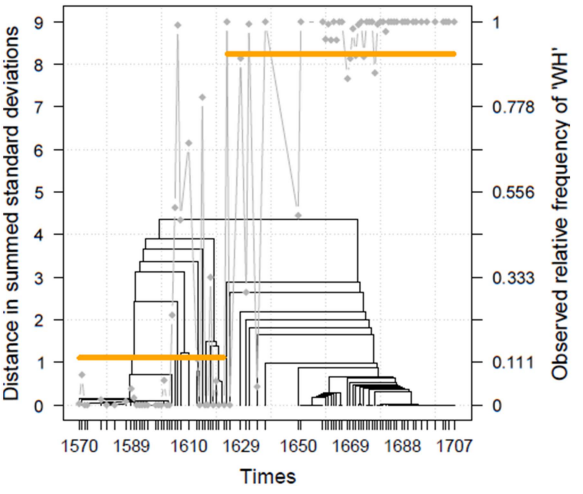

Figure 1. Dendrogram produced by VNC analysis showing change from <quh-> to <wh-> over time

variability, but also similarity in time. The two main clusters in this data are highlighted in yellow. Finally, the graph is overlaid with raw data – each dot represents the frequency of <wh-> within a single text in the HCOS and this is charted on the z-axis. The pattern shown in the graph here is largely what we would expect from previous descriptions of *quh* ~ *wh* variation. There are minor differences in standard deviation initially due to relative uniformity in choice of variant. However, the standard deviations quickly increase in size during the seventeeth century indicating the period of the most variability. Finally, within the last thirty years, there is a drop in standard deviation again and levels return to the pre-1589 levels, suggesting categoricity has been more or less achieved. One thing to notice from the raw data is how little variability exists within individual texts. Even during the seventeeth century, at the height of the change from <quh-> to <wh->, it appears that there is very little intratextual variation. The two clusters found by the VNC analysis (cluster 1: 1570-1623; cluster 2: 1624-1708) indicate that there was an almost binary switch from using <quh-> to <wh-> over a fairly short period of time (1600-50). Individuals, in general, were exhibiting near categorical use of either <quh-> or <wh->.

This dataset contains far fewer clusters than we perhaps might expect of a language change in progress, and certainly fewer than the VNC analysis undertaken by Gries & Hilpert (2010) for *-(e)th* in the *Parsed Corpus of Early English Correspondence* (Nurmi *et al.* 2006). However, Gries & Hilpert (2010) examined a variable that underwent gradual and inconsistent variation over the course of two centuries. The change from <quh-> to <wh-> in Scots on the other hand reflects the artificial and socially conditioned imposition of one language standard over an emerging one, within a few decades of the Union. The period of instability was short-lived and there

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

is little evidence for several periods of rapid then slow change (Meurman-Solin 1993). Rather, the model suggests there was a single, rapid switch from <quh-> to <wh->.

3.2 *Multiple regression*

Next, binomial mixed-effect models were fit to the data by hand (see Baayen *et al.* 2008) using the lme4 package (version 1.1-10; Bates *et al.* 2015) in R with the bobyqa optimizer (Powell 2009) to aid model convergence issues. This was done to determine which extralinguistic factors played a role in driving the change to <wh->. The dependent variable was a binomial variable distinguishing between the spellings <wh-> and <quh-> (coded as either *quh-* or *wh-*). <wh-> was set to the default as the present standard variant and so the data presented here show the log-odds of the <wh-> form. The fixed effects initially coded as independent predictors of variation between <quh-> and <wh-> are presented in table 1.

Some of the levels within the predictor variables presented in table 1 are self-explanatory (e.g. whether the text was from the *Central* or *Northern* region of Scotland). But others require a little more explanation.

AUDIENCE was grouped into six categories: *Documentary (Administrative)*, *Documentary (Public)*, *Public*, *Professional*, *Family* and *Royal/Official. Documentary (Administrative)* refers to texts that were factual rather than imaginative, intended only to be read by the people involved in the transaction. These texts included local records and Acts of Parliament. *Documentary (Public)* also refers to non-imaginative texts though these were intended for or available to the public. In the corpus they consist entirely of histories and trial proceedings. *Public* texts cover a range of text types that could be instructive, fictional or argumentative, but not informational as the documentary texts are. These include travelogues, sermons, pamphlets and handbooks. *Professional* refers to academic literature, in this case scientific and medical treatises. *Family* refers to letters and correspondence addressed to family members, and *Royal/Official* refers to correspondence between members of the Scottish gentry and between the gentry and royal family.

CONTEMPORANEITY also needs some clarification. This was grouped into five categories: *argumentative*, *instructive*, *expository*, *narrative non-imaginative* and *statutory*. *Argumentative* texts were literature involving some form of debate or discussion such as trial procedings or pamphlets. *Instruction* refers to texts intended as guides or directives but with specific audiences in mind. These audiences were either the faithful or royalty, for whom guides were produced concerning correct religious or princely behaviour. *Expository* texts refer to infomative texts such as scientific treatises or handbooks, intended to enlighten the audience on a particular topic. *Narrative non-imaginative* texts is by far the largest category in the corpus, and refers to all non-fiction texts that involve an element of time, including history books, private diaries, bio- and autobiographies (labelled for our purposes as Personal Account) and travelogues. Finally, *Statutory* refers to all texts with a legal element, such as local records and law treatises.

Within LITERARY MEDIUM there are three categories: *written*, *speech-based* and *script*. *Written* refers to the vast majority of documents in the HCOS and encompasses a wide

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

Table 1. *Predictors of variation included in the statistical model predicting variation in <wh-> in Scots between 1570 and 1708*

Predictor variables	Levels
Year	1570–1708
HCOS year groupings	1570–1640
	1640–1708
VNC year groupings	VNC 1 (1570–1623)
	VNC 2 (1624–1708)
Audience	Documentary (administrative)
	Documentary (public)
	Public
	Professional
	Family
	Royal/official
Contemporaneity	Argumentative
	Instruction
	Expository
	Narrative non-imaginative
	Statutory
Text type	Law
	Local record
	Trial proceeding
	Personal account
	Private letter/diary entry
	Letter – non-private
	Pamphlet
	Handbook
	Educational treatise
	History
	Sermon
Text medium	Manuscript
	Printed
Literary medium	Written
	Speech based
	Script
Geographical region	Central Scots
	Northern Scots

range of text types. *Speech-based* refers mostly to trials proceedings in which the defendant is (supposedly) directly quoted, and church proceedings. *Script* refers to religious sermons spoken by preachers to their local congregations. These would have been written in the style of a speech or address, to be delivered in church to a lay audience.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

First, a series of models were generated to explore how best to model time as a predictor in these models. Four simple mixed-effect logistic regression models were run, each having as a predictor one particular method of partitioning time: (i) year as a linear predictor; (ii) the HCOS time periods; (iii) year as a non-linear variable; and (iv) the VNC method of chunking the data over time. All four of these models failed to converge and this is most likely because of the uneven spread of the data over time (i.e. the data are not distributed evenly over each year because we are dealing with historical data; there will be some years with no data points, and others with very many data points). Attempting to build models with time (any of the four methods described above) as a predictor also led to model convergence issues. Since we know that the change from <quh-> to <wh-> took place during this time period, it is less important to incorporate some measure of time as a predictor of variation, and so we removed any measure of time in order to achieve a better statistical model of the data, and one which allows us to explore the social predictors of variation more easily.

Next, before continuing with model fitting, we checked for collinearity between the predictor variables using the vif.mer function in R.[3] Perhaps unsurprisingly, TEXT TYPE, CONTEMPORANEITY (style) and AUDIENCE were all highly correlated. Three models were created with each of the collinear factors entered as the only predictor of variation (i.e. one model explored the extent to which TEXT TYPE was a predictor of <wh-> in the HCOS, another model looked at CONTEMPORANEITY, and other correlated <wh-> with AUDIENCE). χ^2 likelihood tests compared Akaike information criterion (or AIC; Akaike 1974) and Bayesian information criterion (or BIC; Schwarz 1978) values for each of these models. There were no significant differences between these models, but the AIC/BIC values were marginally lower for the model with AUDIENCE so this was selected as the best fixed effect for the social factor addressing style/audience design.

The model was then incrementally expanded to include more extralinguistic factors. At each stage of expansion, the AIC and BIC values were compared with previous models using χ^2 likelihood tests to determine whether the fit of the model was improving. This continued until the model failed to find any more significant predictors of variation. Model convergence issues were encountered when two-way interactions were tested so only the fixed effects are presented here. Random intercepts for AUTHOR and WORD were included, as well as for EDITOR (since for the vast majority of texts there is another potential source of random variability that we can account for by including editor as a random effect).[4] By-author random slopes were checked but

[3] Following the steps outlined here: https://hlplab.wordpress.com/2011/02/24/diagnosing-collinearity-in-lme4/

[4] We chose to include author as a random effect as this follows standard sociolinguistic practice for statistical modelling. In our case we are dealing with written rather than spoken data, and so instead of including speaker as a random effect (as in contemporary sociolinguistic studies), we included author. This assigns a certain amount of variation to the author, enabling the model to take into account that some individuals might vary in ways above or below what the other factors might predict (Johnson 2009). This enables the results to be applicable to the wider population rather than just the subset of authors sampled (see Johnson 2009 for further discussion).

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

Table 2. *Logistic mixed effects regression model of factors predicting the use of <wh-> in Scots between 1570 and 1708 (N = 7,759)*

Random effects	Variance	Standard deviation			
Word	97.99	9.899			
Author	703.35	26.521			
Editor	238.13	15.432			
Fixed effects	**Estimate**	**Std error**	**z-value**	**Pr(>|z|)**	**significance**
(Intercept)	–7.598	8.517	–0.892	0.372317	
Audience description: Documentary/Public	1.581	4.918	0.322	0.74779	
Audience description: Family	25.204	6.395	3.941	8.11E-05	***
Audience description: Professional	7.246	7.404	0.979	0.327788	
Audience description: Public	29.211	7.434	3.929	8.52E-05	***
Audience description: Royal/ Official	2.414	6.726	0.359	0.719718	
Written or spoken: speech-based	12.982	8.083	1.606	0.108228	
Written or spoken: written	28.904	8.574	3.371	0.000749	***
Edited: Yes	–35.988	8.907	–4.041	5.33E-05	***

could not be included because this led to further model convergence issues. The final statistical model showing the predictors to significantly constrain variation in <quh->/<wh-> is reported in table 2.

The estimate is the coefficient estimated by the model – this measures the strength and direction of the relationship between two variables, using the intercept as a reference. The standard error is the error of the estimate, and the z-value is a measure of standard deviation (thus measures closer to zero will be closer to the mean – in this case the intercept). The probability value Pr(>|z|) reports the likelihood that the correlation between the dependent variable and predictor variable is due to chance.

It is clear that the anglicised variant <wh-> is significantly affected by three factors, reflecting the interweaving influences of extralinguistic constraints. The various AUDIENCE categories also exhibit considerable variability, suggesting AUDIENCE had a great influence on the variation observed for <wh->. Of course, the effects presented here are unlikely to be the only significant features; author-specific characteristics such as gender or rank may have been important too but the data for these social characteristics are patchy in the HCOS so it wasn't possible to include these as predictors. We now discuss each of these significant constraints on the rise of <wh-> in Scots.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

4 Results and discussion

4.1 Audience

AUDIENCE reflected substantially more variation across its factor levels than contemporaneity, supporting its validity as a conditioning factor of anglicisation. Figure 2 depicts the likelihood of the variable being realised as <wh-> (y-axis) across different audience types (labelled on the x-axis). The higher the value on the y-axis, the more likely that the variant would be realised as the incoming, anglicised <wh-> form. There is a clear cline in AUDIENCE from a near-categorical preference for Scots <quh-> to an increasing degree of <wh-> forms, suggesting different audiences encouraged or constrained the use of <wh->. However, anglicisation was clearly strongest in texts addressed to the public.

The preference for the anglicised variant in the *Public* category is hardly surprising. Meurman-Solin (1993) has suggested that authors of such texts may have been motivated to reach a wider audience, to enjoy the benefits offered by 'high' society whilst also maintaining the profitability of printing. The Union evidently increased opportunities for social advancement and focused the gentry's attention away from Edinbugh towards London (Lawson 2014). Parties were no longer concerned solely with their Scots audience but also with readers outside Scotland's borders. Public access could best be achieved through use of the incoming standard, allowing authors greater scope than if they restricted themselves to purely Scots forms and styles (Meurman-Solin 1993). Thus, texts aimed at the public can be expected to incorporate anglicised forms the most, and this is indeed the case.

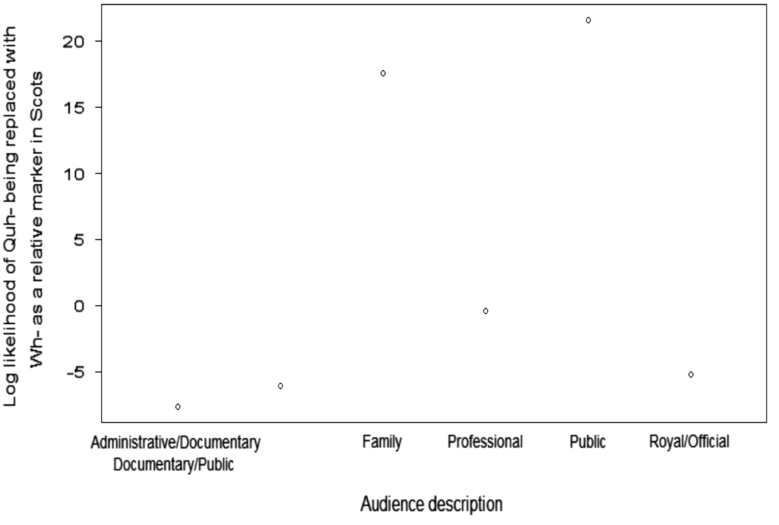

Figure 2. Model output showing likelihood of <wh-> across different audience categories in the *Helsinki Corpus of Older Scots* (1570–1708)

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

It is equally unsurprising to find that texts coded as *Administrative/Documentary* are the least anglicised out of all the audience types. This category has the strongest negative correlation with *Public*, and figure 2 implies almost zero deviance from the standard Scots <quh-> form. This result is hardly unexpected given that such texts were never intended to be read by the public, but rather by various bureaucratic officials who did not need to be influenced, persuaded or appealed to in any way. Meurman-Solin (1992, 1994) has noted that genres with no particular addressee tend to remain linguistically conservative and this is reflected here. Furthermore, Scots differed from English in the legal arena in that it had a different professional terminology. By following set conventions, scribes guaranteed the reliability and transparency that records required (Bugaj 2004, 2005; Kopaczyk 2012, 2013; Cruickshank 2013). This included the prepositional phrase fragment *witness of ye quhilk/quhilkis* ((the) witness of (the) which), which could possibly explain the extended life of the <quh-> spelling in legal discourse (Kopaczyk 2013). Indeed, Meurman-Solin (1989c) found no definite change in trials and law, with <quh-> variants resisting anglicising tendencies longer than other features. These codified expressions may be part of the reason for its retention, but without economic and social pressure to convert to anglicised forms, it seems unlikely that there was much appeal to do so regardless.

Texts directed at a *Professional* audience, which consisted of scientific and medical texts, are more anglicised than *Administrative/Documentary* but still largely conservative. This may be indicative of the changing demands of the professional audience during this time. Scotland had a relatively long-established scholarship that was recognised beyond its borders. Certainly during the fourteenth and fifteenth centuries Scotland produced a great deal of leading work in various fields (Bald 1926). Unlike English scholarship, which was conducted in Latin or French, Scottish scholars had already begun to publish some works in the vernacular earlier on, especially when publishing for a wider audience (Bugaj 2004, 2005). These processes would have inhibited a large-scale influx of anglicised forms initially. Yet it appears anglicisation did catch on; as England began to produce scientific literature in the vernacular, new expectations and preferences regarding the language of scholarship were formed. Scots scholars and scientists seeking to publish their work for the wider academic community would have been pushed to employ an anglicised form, rather than Scots. Nevertheless, it appears these changes were somewhat (though not significantly) slower to reach completion than the changes occurring to texts intended for the public. Hence, we see an increased level of <wh-> for this audience type relative to *Administrative/Documentary* texts.

It is quite surprising, however, to find texts addressed to the *Family* to be the second most anglicised, whilst *Royal/Official* texts are not particularly anglicised at all. This would seem to contradict our expectations; the familiarity, intimacy and codified conventions of personal communication would predict the continued use of Scots forms, whilst the London-based monarchy would be expected to encourage anglicisation. Texts falling under the *Family* category consist largely of letters sent back and forth between the gentry in London and their family members back home. The landed gentry spent an increasing part of their time in London following the Union, in order

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

to partake in the new social opportunities that were formed as a consequence. These surroundings may have influenced their language use accordingly, whilst family members writing to noblemen could hardly fail to be aware of their addressees' shift to the capital of the Southern English standard. The changing social situation may thus account for why these texts are more anglicised than their *Adminstrative/Documentary* counterparts. However, without full demographic and personal information regarding the addressee and author of these texts, such possibilities must remain speculative for now.

The conservative nature of texts addressed to a *Royal/Official* audience is more difficult to explain. This category contains letters and works written to/for King James VI of Scotland (James I of England) as well as works by the monarch himself. Scots was perhaps used with the Scottish king to develop a sense of in-group identity and intimacy in order to gain trust. Despite King James VI's residence in London and his kingship over both Scotland and England, his origins were Scottish and those writing to him could hardly fail to be aware of this. Unlike the wider public, the recipient here (King James VI) was familiar with Scots and thus there was perhaps little need to adopt anglicised forms. Furthermore, James himself was aware of the delicate state of the Scots language and its decline. Indeed, he wrote a treatise urging distinctiveness and championing the vernacular in particular rhetorical situations (Jack 1997). Being in the ultimate position of power, there was also less need for James VI to anglicise in order to move upwards in social circles. Again, however, such explanations are simply possibilities and this could certainly warrant further investigation.

It appears that one audience type in particular accelerated the anglicisation of Scots: *Public*. To some extent this may be because other audience types had a fixed format that was more or less constant, regardless of exterior political and social changes. The wider public, however, was fluid. This was not a fixed set of individuals but a constantly shifting norm that changed with the times, and at a rapid pace. The Union of 1603 increased the audience pool dramatically and thus change was not only preferred, it was necessary. However, this is not to say that all Scots people unanimously adopted English once the Union was complete. Indeed, some viewed the anglicising trend as profoundly unpatriotic and distasteful (Jones 1997a; Cruickshank 2013). The majority, however, had little choice if they wished to perpetuate their work beyond a purely Scots audience, and thus pragmatic concerns dictated their writing style. Nevertheless, one must be cautious in interpreting the entire literary development of Scots using only these data. This is simply the path of one orthographic variant across a select number of texts during a particular time in Scots history. It is too simplistic to suggest that <quh->/<wh-> variation can act as a proxy for the displacement of Scots by the new incoming standard. At most, it is suggestive of wider changes and patterns that were affecting Scots during this time, though each incoming variant may have its own specific path and manifestations.

These data do indicate, however, that it is AUDIENCE rather than TEXT TYPE that perhaps needs to be investigated in more detail. This differs from previous analyses which have simply assumed that TEXT TYPE is the most important predictor of variation (Aitken 1979; Devitt 1989; Meurman-Solin 1989b, 1992, 1993, 1994;

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

Görlach 1998). Although Meurman-Solin (1993) has argued that AUDIENCE was an important influence in the use of anglicised variants, TEXT TYPE has remained the central component of her examinations. Indeed, she has suggested that ultimately AUDIENCE and STYLE describe TEXT TYPE rather than acting independently of it (2003). Yet the results presented here suggest quite the opposite.

Furthermore, relying on TEXT TYPE is problematic given the difficulty in trying to circumscribe individual genres. There is little information available on how text types were understood by their authors in the sixteenth century. It is not clear whether the codified expectations and textual format argued to have influenced certain genres had yet been consolidated, perhaps allowing them a certain level of flexibility that is not always acknowledged in historical analyses. Instead it appears that historical research – at least that examining Scots – needs to begin on a more basic, fundamental level: with the readers of the text who ultimately determined its use and dissemination through society. Indeed, the effect of AUDIENCE over TEXT TYPE and CONTEMPORANEITY is perhaps not so surprising; the audience could well dictate the appropriate style and format of a text to a certain extent, playing the ultimate role in a text's final production.

Despite the value of AUDIENCE as a predictor variable, it is highly unlikely that a single factor drove forward the change, given the number of conditioning factors that can operate on any instance of language change. The mixed-effects regression model also identified LITERARY MEDIUM (*written*, *spoken*, *script*) as a signficant effect that interacted with and drove the changeover to <wh->.

4.2 *Literary medium*

Figure 3 plots the likelihood of <wh-> (y-axis) in the three literary mediums present in the HCOS (x-axis). It is clear that *Written* texts behave significantly differently to

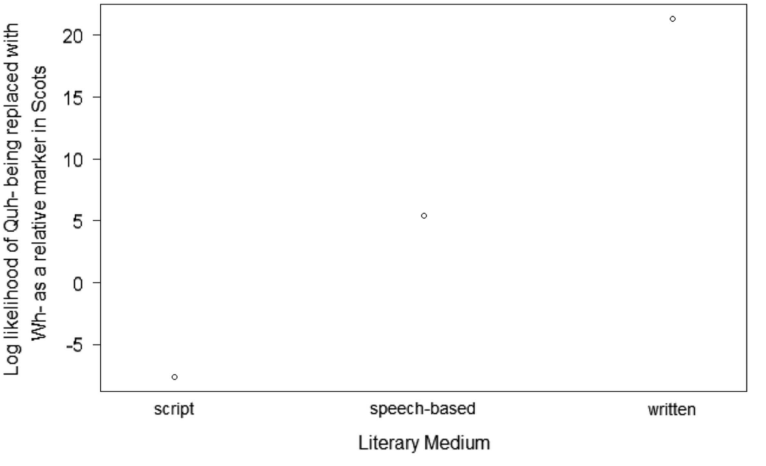

Figure 3. Likelihood of <wh-> across different literary mediums in the *Helsinki Corpus of Older Scots* (1570–1708)

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

Scripts, reflecting a clear preference for <wh->, though both *Written* and *Speech-based* texts exhibit higher levels of the incoming variant than *Script.*

The tendency for *Written* texts to prefer the anglicised variant is not so surprising, given that writing is an inherently conscious act, but the proportion of anglicisation in *Speech-based* texts is unexpected. This runs contrary to earlier research, which has stressed that spoken language was the last to anglicise (MacQueen 1957; Aitken 1979, 1997; Beal 1997). Whilst this result may be suggestive of the changing times in Scots speech, and the underlying phonological processes that affected this variant in Middle English (Lass & Laing 2016), it seems unlikely these processes alone can explain the high levels of <wh-> that occurred so rapidly after the Union. Instead, the result we see here is more likely the result of scribal tendencies, though there may have been influence from sound changes that were already underway at this point. The temporal structuring of the data suggests we are seeing a sudden change in spelling practices rather than a mass convergence in the speech of the Scots. Scribes may have applied their own editing practices, which will remain forever unknown to us, whilst their alteration of the orthographic information did not alter the semantic content of the trial. The scribe noting down the defendent's speech could switch to using <wh-> while still preserving quite faithfully what was said. Despite the perceived trend, however, *Speech-based* texts did not differ greatly from *Scripts*; the mixed-effects model indicated that the relationship between the two was only weakly significant.

The position of *Script* as less anglicised than *Speech-based* texts is also an interesting case. The result here seems to contradict Aitken's (1979) claim that sermons were partly modelled on Biblical English following the Reformation. This also suggests that not all religious writings were equally anglicised after 1560. Indeed, Tulloch (1997) has argued that Scots tended to be preserved in sermons and religious texts that were aimed specifically at Scots audiences, and this appears to be the case here. Despite use of the English Bible and Psalter, preachers may have recognised the local nature of their audience, who would have felt alienated by purely English use in their local parish. Furthermore, these texts were not intended to be seen by anyone other than the preacher. Thus, given that sermons were intended to be read aloud, and there was no particular motivation to anglicise the texts, such considerations might lead the clergy to favour the variant that was orthographically (and possibly phonologically) Scots. Audience is key in explaining this trend, as a textual analysis that categorised sermons under 'religious texts' on the other hand would fail to observe the differences between local sermons and religious treatises that were directed at the wider audience abroad.

4.3 Edited

Finally, the mixed-effects regression model identified EDITED (whether the text was edited or not) as a signficant effect that interacted with and drove the changeover to <wh->. Figure 4 plots the likelihood of <wh-> (y-axis) in *Edited* and *Unedited* texts present in the HCOS (x-axis). There is a clear preference for <quh-> in *Edited* texts whilst *Unedited* texts exhibit a higher level of <wh->.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

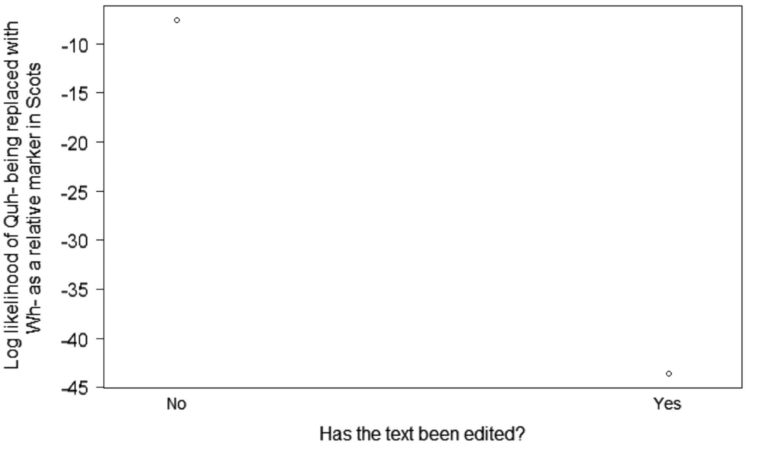

Figure 4. Likelihood of < wh- > across different literary mediums in the *Helsinki Corpus of Older Scots* (1570–1708)

This is a surprising effect as it manifests in opposite ways to what we would expect, and is perhaps counterintuitive. Indeed, earlier studies have suggested that editors would have played a standardising role in Middle Scots, choosing the anglicised variant more often (Devitt 1989: Meurman-Solin 1993, 2003). Meurman-Solin (1993) has claimed that many editors within the corpus modernised spellings, and any changes made were consistently in favour of anglicised variants. Yet relatively little is known about the practices and procedures involved in preparing a text for printing and publication (Meurman-Solin 1993). The role of the editor and the extent of their input will remain forever unknown to us. However, our previous results suggest authors were aware of the orthographic difference, and the presence of the anglicised variant was often an explicit choice made by the author, rather than the product of sub-conscious interference from English. This may explain in part the anglicisation of unedited texts, which could easily have been undertaken by the author themselves.

This does not adequately explain, however, why edited texts are much less anglicised in the corpus, but the selection process behind the texts in the corpus provides a possible clue. Texts in the corpus are split into two basic categories: texts of which the earliest or contemporary printed version has been used, and texts printed in the nineteenth and twentieth centuries, edited from earlier manuscripts (Meurman-Solin 1993). The first group contains edited, but also unedited, texts printed during the time period in question, whereas the second group comprises entirely of edited prints that were produced much later on but 'chosen for their linguistic value' (Meurman-Solin 1993: 140). It may be that texts reflecting a higher prevalence of Scots features were favoured for this second group, as well as a preference for manuscripts produced early on in the change. This could lead to a higher proportion of Scots features within edited texts overall. There is evidence that later editions produced during the seventeenth and eighteenth centuries reflected a more uniform mode of spelling (Meurman-Solin

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

Table 3. *The number and proportion of texts which are edited, compared with those which are not, cross-tabulated with year group (as defined by the VNC analysis; see table 1)*

Year category (as defined by VNC)	Number and proportion of texts in the HCOS which are not edited	Number and proportion of texts in the HCOS which are edited
VNC 1 (1570–1623)	620 (17.3%)	2,956 (82.7%)
VNC 2 (1624–1708)	1667 (39.9%)	2,516 (60.1%)

1993). As English became increasingly codified as the language of print, editors and authors alike often decided to modernise spellings in the direction of the anglicised variants (Meurman-Solin 1993). However, this change did not take off overnight, and thus a selection criteria based on early productions of works could see comparatively low levels of anglicisation. Furthermore, there is a much higher proportion of non-edited texts occurring later in the HCOS, when the switch from <quh-> to <wh-> was already well underway, while there is a higher proportion of edited texts from earlier in the time frame, when <quh-> to <wh-> was still very variable (see table 3).

There are therefore a number of possible explanations for why there is less <wh-> in edited texts than we might expect

4.4 Random intercepts

Finally, it is interesting to consider the role of the individual author producing the texts. Recent work in sociolinguistics has shown the value of inspecting random intercepts from a mixed-effect regression model to explore the role of individual speakers in leading or lagging behind in specific changes in progress (Drager & Hay 2012; Watson & Clark 2016). When predictor variables are included in a regression model as random effects, rather than fixed effects, each level within that predictor (e.g. each author in this case) is assigned a value (called the random intercept) and the value is calculated based on how much unexplained variance there is for that level (i.e. that author) in the model. Drager & Hay (2012) showed that, in a corpus of speech, individual speakers with the lowest intercepts were those that were not adopting the innovation and those with the highest intercepts were leading the speech community towards the new variants. These effects are over and above those that are reported in the model as main effects (in other words, exploring the random intercepts does nothing to change the main effects within our model: AUDIENCE, LITERARY MEDIUM and EDITED texts are still significantly influencing the shift from <quh-> to <wh->). With this in mind, we can explore the random effects to find out which of these authors were using more or less <wh-> than expected. The random intercepts for *Author* are plotted in figure 5.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

The line down the centre of figure 5 signals the position of zero. Individual authors who fall somewhere along that line are not using more or less of <wh-> than would be expected from the model, once other predictors have been accounted for. Those with values to the left of the line are using less <wh-> than expected and those with values to the right of the line are using more <wh-> than would be expected for them.

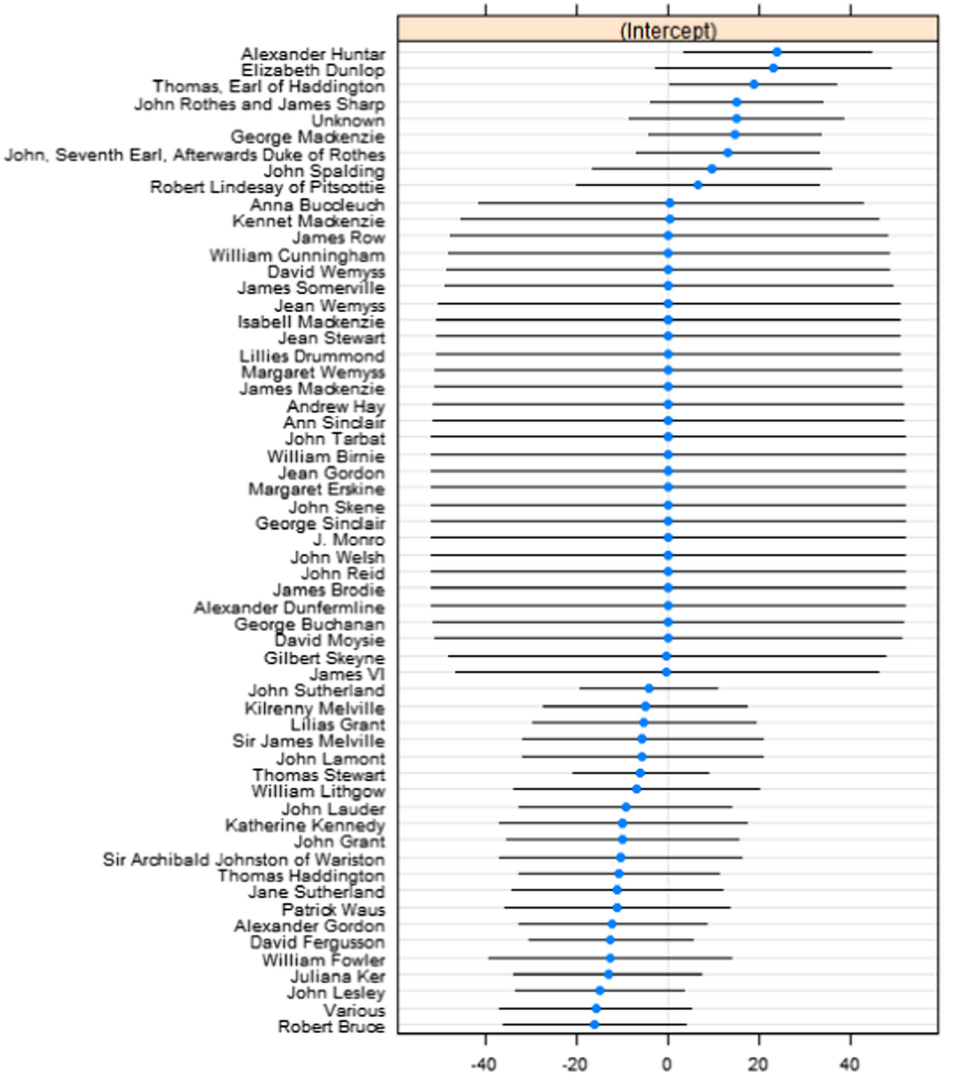

Figure 5. Random intercepts for Author where known in all texts from the *Helsinki Corpus of Older Scots* (1570–1708). Multiple authors are marked as 'various' and texts where the author was anonymous are marked as 'unknown'

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

A brief examination of some of these authors yields interesting case studies. Thomas Hamilton, First Earl of Haddington (1563–1637), reflects innovative language use. He was on very good terms with James VI and also served Charles I. He was appointed to manage the finances of Scotland, and later took on influential roles in the Privy Council and government. Following the Union, English quickly became the common language of the kingdom and monarchy, thus his connections with the Parliament and monarchy may explain why he chose to adopt the incoming standard at a greater rate than expected. George Mackenzie (1636–91) also exhibits higher-than-expected use of <wh->. Mackenzie was a lawyer and Lord Advocate, and a member of the Scottish Parliament and Privy Council of Scotland. He was the minister responsible for the persecuting policy of Charles II in Scotland against the Presbyterian Covenanters, and also opposed the dethronement of James II. Again, his royalist loyalties are clear and this commitment to the monarchy and the unified kingdoms of England and Scotland may have encouraged use of the national standard. In addition, Mackenzie did not share the separatist, autonomous opinions of his Covenanting contemporaries or have any qualms about persecuting his fellow countrymen. Furthermore, Mackenzie and Haddington, through their careers, would have had greater exposure to the English standard given their proximity to its centre of influence. Such factors (above and beyond those already found in the model) may be responsible for why we see a greater use of <wh-> in the linguistic repertoire of these individuals.

On the other hand, Lord Archibald Johnston of Wariston (1611–63) reflects conservative language usage. He took a major role in writing the Scottish National Covenant in 1638, which effectively undermined the established church supported by the monarchy, and was a key facilitator in negotiating the peace treaties of Berwick in 1639 and Ripon in 1640. These treaties, ending the first and the second Bishops' Wars between England and Scotland, were a humiliating defeat for King Charles I, who had to make considerable concessions to Scotland as a result. Johnston opposed royal intervention in Scottish affairs, particularly regarding the ecclesiastical structure of Scotland, and publicly spoke out against royal proclamations regarding the Church and Parliament. He opposed monarchical control of state appointments and probably drew up the Act of Classes (1649) which banned royalists from holding public office in Scotland. Johnston was clearly anti-royalist, nationalistic and firmly focused on Scotland and its right to maintain its ancient legal and ecclesiastical structures. This political ideology could explain why his use of <quh-> is higher than expected, particularly given that Johnston was writing after the switch had largely taken place. Interestingly, Johnston's political life reflects the exact opposite of Mackenzie; he supported the very movement Mackenzie sought to crush, and their language usage is similarly contrary. William Fowler (1560–1612) is another individual who is more conservative than the model would predict. Fowler was a makar (a royal bard), writer, courtier and translator, becoming part of a literary circle around King James VI. Fowler produced poetry, sonnets, treatises and pamphlets that were commended by the king himself. Scottish vernacular literature was one of few literary arenas less influenced by the prestige variety of the South, and Scots features could persist far longer

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

in such works than in other written domains. Thus, Scots features were permitted in Fowler's works despite being intended for the public.

Again, it is difficult to say with absolute certainty which factors influenced the idiosyncrasies of particular historical actors. However, by exploring random intercepts we are able to see which individuals are leading the change and who lags behind, indicating interesting trends and the highly individual nature that language change can assume once the analysis is broken down to the micro-level.

One final point to notice about the results presented here is that there was no significant effect for printed texts. Printing has been argued by various scholars (Bald 1926, 1927; MacQueen 1983; Meurman-Solin 1993; Kniezsa 1997) to have had an influential role in anglicising Scottish works, yet the model failed to find a significant difference between handwritten manuscripts and those that were printed. This may suggest that printing had little effect on the <wh-> variant, or that the significance of AUDIENCE is so great that it overrides a discernible difference between the two textual mediums. This is something that would warrant further investigation.

4.5 Overview

In summary, the trends found through mixed-effects modelling (rather than purely descriptive statistics) highlight the value of analysing multiple competing influences operating on the rise of <wh->. Previous analyses of <quh-> ~ <wh-> in the HCOS have perhaps overestimated the effect of TEXT TYPE and PRINTING on the processes of anglicisation. Yet, by utilising mixed-effects modelling, relationships that would otherwise remain hidden within the larger framework of historical literature can be uncovered, validated and linked to identifiable sociohistorical changes. Of course we understand that this is a study of a single variant and its patterning in the decline of written Scots and rise of SSE. In order to confirm whether the trajectory and the significant effects identified here hold for the anglicisation of Scots during 1570–1708 in general, we need to explore more linguistic variables. This might indicate which factors were influential across the board for anglicisation, and which were specific to different variants. Such an analysis would also indicate whether most Scots variants underwent a sudden, binary switch or whether some were more prone to variable use by the same author. The latter might indicate Scots variants that were perhaps linguistically or orthographically less salient. Alongside this, a study examining SSE beyond 1708 would also be enlightening. This could indicate whether there were variants that took longer to reach categoricity, or whether other factors became more important in conditioning the variation later on.

5 Concluding remarks

Our main contribution in this article is to show that by incorporating modern statistical methods that are used frequently in the analysis of contemporary corpus data (Hay *et al.* 2015; Gries 2016) we can reach a clearer understanding of the factors which drove language change in the history of Scots. In historical linguistics generally, we

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

understand that this is not new – work on the history of English has employed these techniques for some time (Nevalainen 2006; Nevalainen & Raumolin-Brunberg 2003; Hinneburg *et al.* 2007; Gries & Hilpert 2010), but they have been slow to catch on in work on Scots. Previous accounts of language change have continued to rely on descriptive statistics. We hope to have shown that by allowing the data to demonstrate the importance of certain social factors, rather than arbitrarily imposing the focus of the investigation or presupposing the importance of a social factor, we have presented a different picture of this instance of historical language change. Specifically, while previous work has pinpointed TEXT TYPE as central to the shift from <quh-> to <wh-> in Scots, our work shows that in fact it is AUDIENCE that seems to have been a more impotrant predictor of variation.

Furthermore, we follow Gries & Hilpert (2008, 2010, 2012) in calling for scholars of the history of Scots (as they have for scholars of the history of English) to avoid sectioning historical data into convenient time periods as this can disguise or overlook trends in the trajectory of language change. Our work suggests that the shift to the anglicised variant occurred rapidly and was a largely binary switch in the minds of most authors. This indicates that the change was not the result of a gradual process of natural language change or increasing pressure from England over time, but the sudden, artificial imposition of one emerging standard over another.

Of course, we are presented with a finite snapshot of the past in limiting ourselves to one corpus and one variable; 'corpora are always incomplete models of some linguistic reality' (Gries & Hilpert 2010: 297). Unfortunately, historical data will always be limited in this way. Nonetheless, modern, statistical analyses can bring us as close as currently possible to a more thorough understanding of underlying diachronic developments, and their manifestation in a particular variety at a particular time.

Authors' address:

Department of Linguistics
University of Canterbury
Private Bag 4800
Christchurch 8140
New Zealand
sarah.ve@outlook.com
lynn.clark@canterbury.ac.nz

References

Aitken, Adam Jack. 1979. Scottish speech: A historical view with special reference to the Standard English of Scotland. In Aitken & McArthur (eds.), 85–120.

Aitken, Adam Jack. 1984. Scots and English in Scotland. In Peter Trudgill (ed.), *Language in the British Isles*, 517–32. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

Aitken, Adam Jack. 1997. The pioneers of anglicised speech in Scotland: A second look. *Scottish Language* 16, 1–36.
Aitken, Adam Jack & Tom McArthur (eds.). 1979. *Languages of Scotland*. Edinburgh: W&R Chambers.
Akaike, Hirotugo. 1974. A new look at the statistical model identification. *Transactions on Automatic Control* 19(6), 716–23.
Anthony, Lawrence. 2015. *AntConc (version 3.5.0)* [computer software]. Tokyo: Waseda University. www.laurenceanthony.net
Baayen, R. Harald, Douglas J. Davidson & Douglas M. Bates. 2008. Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4), 390–412.
Bald, Marjory. 1926. The anglicisation of Scottish printing. *The Scottish Historical Review* 23 (90), 107–5.
Bald, Marjory. 1927. The pioneers of anglicised speech in Scotland. *The Scottish Historical Review* 24(95), 179–93.
Bates, Douglas, Mächler Maechler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48. doi:10.18637/jss.v067.i01
Beal, Joan. 1997. Syntax and morphology. In Jones (ed.), 335–77.
Blake, Norman F. 1992. Translation and the history of English. In Rissanen *et al*. (eds.), 3–24.
Bugaj, Joanna. 2004. Middle Scots as an emerging standard and why it did not make it. *Scottish Language* 23, 19–34.
Bugaj, Joanna. 2005. Middle Scots burgh court records: The influence of the text type on its linguistic features. In Nikolaus Ritt & Herbert Schendl (eds.), *Rethinking Middle English: Linguistic and literary approaches*, 75–88. Frankfurt am Main: Peter Lang.
Clark, Lynn & Kevin Watson. 2016. Phonological levelling, diffusion, and divergence: /t/ lenition in Liverpool and its hinterland. *Language Variation and Change* 28(1), 31–62.
Cruickshank, Janet. 2013. The role of communities of practice in the emergence of Scottish Standard English. In Joanna Kopaczyk & Andreas H. Jucker (eds.), *Communities of practice in the history of English*, 19–45. Amsterdam: John Benjamins.
Devitt, Amy. 1989. *Standardising written English: Diffusion in the case of Scotland, 1520–1659*. Cambridge: Cambridge University Press.
Dictionary of the Scots Language / Dictionair o the Scots Leid. 2004. Scottish Language Dictionaries Ltd. www.dsl.ac.uk (accessed 16 August 2015).
Douglas, Sheila. 2001. Scots language and the song tradition. In John Monfries Kirk & Dónall Ó. Baoill (eds.), *Language links: The languages of Scotland and Ireland*, 233–6. Belfast: Cló Ollscoil na Banríona.
Drager, Katie & Jennifer B. Hay. 2012. Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change* 24, 59–78.
Görlach, Manfred. 1996. And is it English? *English World-Wide* 17(2), 153–74.
Görlach, Manfred. 1998. Text types and the hstory of Scots. In *Even more Englishes: Studies 1996–1997* (Varieties of English around the World G22), 55–77. Amsterdam: John Benjamins.
Gries, Stefan Th. 2016. *Quantitative corpus linguistics with R*, 2nd rev. and ext. edn. London and New York: Routledge.
Gries, Stephan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based neighbor clustering. *Corpora* 3(1), 59–81.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

Gries, Stephan Th. & Martin Hilpert. 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, Stephan Th. & Martin Hilpert. 2012. Variability-based neighbour clustering: A bottom-up approach to periodization in historical linguistics. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford handbook on the history of English*, 134–44. Oxford: Oxford University Press.
Hay, Jennifer B., Janet B. Pierrehumbert, Abby J. Walker & Patrick LaShell. 2015. Tracking word frequency effects through 130 years of sound change. *Cognition* 139, 83–91.
Hinneburg, Alexander, Heikki Mannila, Samuli Kaislaniemi, Terttu Nevalainen & Helena Raumolin-Brunberg. 2007. How to handle small samples: Bootstrap and Bayesian methods in the analysis of linguistic change. *Literary and Linguistic Computing* 22(2), 137–50.
Jack, Ronald D. S. 1997. The language of literary materials: Origins to 1700. In Jones (ed.), 213–63.
Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3(1), 359–83.
Jones, Charles. 1997a. Introduction. In Jones (ed.), 1–5.
Jones, Charles (ed.). 1997b. *The Edinburgh history of Scots*. Edinburgh: Edinburgh University Press.
King, Anne. 1997. The Inflectional Morphology of Older Scots. In Jones (ed.), 156–83.
Kniezsa, Veronika. 1997. The origins of Scots orthography. In Jones (ed.), 24–46.
Kopaczyk, Joanna. 2012. Communication gaps in seventeenth century Britain: Explaining legal Scots to English practitioners. In Barbara Kryk-Kastovsky (ed.), *Intercultural miscommunication past and present* (Warsaw Studies in English Language and Literature), 217–43. Berlin: Peter Lang.
Kopaczyk, Joanna. 2013. How a community of practice creates a text community: Middle Scots legal and administrative discourse. In Joanna Kopaczyk & Andreas H. Jucker (eds.), *Communities of practice in the history of English*, 225–47. Amsterdam: John Benjamins.
Laing, Margaret & Roger Lass . Forthcoming. Old and Middle English spellings for OE *hw*-, with special reference to the 'qu-' type: In celebration of LAEME, (e)LALME, LAOS and CoNE. In Rhona Alcorn, Bettelou Los, Joanna Kopaczyk & Benjamin Molineaux (eds.), *Historical dialectology in the digital age*. Edinburgh: Edinburgh University Press.
Lass, Roger & Margaret Laing. 2016. Q is for what, when, where? The 'q' spellings for OE hw-. *Folia Linguistica Historica* 37, 61–110.
Lawson, Robert. 2014. *Sociolinguistics in Scotland.* New York: Palgrave Macmillan.
MacQueen, Lilian Edith Cochrane. 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. PhD thesis, University of Edinburgh. www.era.lib.ed.ac.uk/handle/1842/7316
MacQueen, Lilian Edith Cochrane. 1983. English was to them a foreign tongue. *Scottish Language* 2, 49–51.
McClure, J. D. 1983. *Scotland and the Lowland tongue*. Aberdeen: Aberdeen University Press.
Meurman-Solin, Anneli. 1989a. The Helsinki Corpus of Older Scots. In Meurman-Solin (ed.), 218–26.
Meurman-Solin, Anneli. 1989b. Variation analysis and diachronic studies of lexical borrowing. In Graham D. Caie (ed.), *Proceedings of the Fourth Nordic Conference for English Studies*, 1, 87–98. Copenhagen: Department of English, University of Copenhagen.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

Meurman-Solin, Anneli. 1989c. Variation and variety in Middle Scots reconsidered: A test study of the Helsinki Corpus of Older Scots. In Meurman-Solin (ed.), 236–46.
Meurman-Solin, Anneli. 1992. On the morphology of verbs in Middle Scots: Present and present perfect indicative. In Rissanen *et al.* (eds.), 611–23.
Meurman-Solin, Anneli. 1993. *Variation and change in early Scottish prose: Studies based on the Helsinki Corpus of Older Scots*. Helsinki: Suomalainen Tiedeakatemia.
Meurman-Solin, Anneli. 1994. On the evolution of prose genres in Older Scots. *Nowele* 23, 91–138.
Meurman-Solin, Anneli. 1995. *The Helsinki Corpus of Older Scots* www.helsinki.fi/varieng/CoRD/corpora/HCOS/ (accessed 23 March 2015).
Meurman-Solin, Anneli. 1997. Differentiation and standardisation in Early Scots. In Jones (ed.), 3–23.
Meurman-Solin, Anneli. 2003. Corpus-based study of Older Scots grammar and lexis. In Jeremy Corbett, J. D. McClure & Jane Stuart-Smith (eds.), *The Edinburgh companion to Scots*, 170–96. Edinburgh: Edinburgh University Press.
Millar, Robert McColl. 2005. *Language, nation and power: An introduction*. Basingstoke: Palgrave Macmillan.
Murison, David. 1979. The historical background. In Aitken & McArthur (eds.), 1–13.
Nevalainen, Terttu. 2006. Historical sociolinguistics and language change. In Ans van Kemenade & Bettelou Los (eds.), *The handbook of the history of English*, 558–82. Oxford: Blackwell.
Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England*. London: Routledge.
Nurmi, Arja, Ann Taylor, Anthony Warner, Susan Pintzuk & Terttu Nevalainen. 2006. *Parsed Corpus of Early English Correspondence*, tagged version. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.
Pollner, Clausdirk. 2000. Shibboleths galore: The treatment of Irish and Scottish English in histories of the English language. In Dieter Kastovsky & Arthur Mettinger (eds.), *The history of English in a social context: A contribution to historical sociolinguistics*, 363–76. Berlin: Mouton de Gruyter.
Powell, M. J. 2009. The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06*, 26–46. Cambridge: University of Cambridge.
Rissanen, Matti, Merja Kytö, Leena Kahlas-Tarkka, Matti Kilpiö, Saara Nevanlinna, Irma Taavitsainen, Terttu Nevalainen & Helena Raumolin-Brunberg. 1991. *The Helsinki Corpus of English Texts*. www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/
Rissanen, Matti, Ossi Ihalainen, Terttu Nevalainen & Irma Taavitsainen (eds.). 1992. *History of Englishes: New methods and interpretations in historical linguistics*. Berlin: Mouton de Gruyter.
Romaine, Suzanne. 1982. *Socio-historical linguistics: Its status and methodology*. Cambridge: Cambridge University Press.
Sankoff, D. 1975. *VARBRUL 2*. Unpublished program and documentation.
Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–4.
Smith, Daisy. Forthcoming. The predictability of {S} abbreviation in Older Scots manuscripts. In Rhona Alcorn, Bettelou Los, Joanna Kopaczyk & Benjamin Molineaux (eds.), *Historical dialectology in the digital age*. Edinburgh: Edinburgh University Press.
Tulloch, Graham. 1997. Lexis. In Jones (ed.), 378–435.

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press

APPENDIX

Table A1. *The original category labels of the* Helsinki Corpus of Older Scots, *and the new categories after regrouping of similar levels*

Corpus variable	Pre-existing categories	Which categories were merged	Final categories for analysis
Audience design	Family	Family +	Family
	Documentary/Family	Documentary/Family	Royal/Official
	Royal	Royal + Royal/Public +	Public
	Royal/Public	Official	Professional
	Official	Professional/Public +	Administrative/ Documentary
	Professional/Public	Public	Documentary/Public
	Professional	Professional/Public +	
	Public	Professional	
	Documentary Public		
	Administrative/ Documentary		
Contemporaneity	Argumentative	Argumentative +	Argumentative
	Argumentative/ Narration	Argumentative/Narration	Expository
	Non-Imaginative	Non-Imaginative	Instruction
	Expository	Instruction Religious +	Narrative Non-Imaginative
	Instruction Religious	Instruction Secular	Statutory
	Instruction Secular		
	Narrative Non-Imaginative		
	Statutory		
Text type	Autobiography	Autobiography +	Diary/Letter Private
	Biography Other	Biography Other +	Educational Treatise
	Diary Private	Travelogue	Handbook
	Educational Treatise	Diary Private + Letter	History
	Handbook Other	Private	Law
	History	Handbook Other +	Letter Non-Private
	Law	Science Medicine +	Local Record
	Letter Non-Private	Science Other	Pamphlet
	Letter Private		Personal Account
	Local Record		Proceeding Trial
	Pamphlet		Sermon
	Proceeding Trial		
	Science Medicine		
	Science Other		
	Sermon		
	Travelogue		

https://doi.org/10.1017/S1360674319000078 Published online by Cambridge University Press