THE UNIVERSITY
of EDINBURGH

ye saidſ lettrez:
The orthographic representation of inflectional morphemes in Older Scots

Daisy Smith

A thesis submitted for the degree of Doctor of Philosophy
University of Edinburgh

2018

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Parts of chapter 1 and chapter 10 of this work are adapted from a work which has been published as follows:

Smith, D. (2019). "The Predictability of {S} Abbreviation in Older Scots Manuscripts According to Stem-final *Littera*". In: *Historical dialectology in the digital age*. Ed. by Alcorn, R., Kopaczyk, J., Los, B., and Molineaux, B. Edinburgh: Edinburgh University Press.

Daisy Smith

Signature Date

Abstract

The general tendencies characterising Older Scots (OSc) inflectional morphology and differentiating it from that of Middle English (ME) have been described (Minkova 1991; King 1997; Aitken 1977; Aitken and Macafee 2002; Kopaczyk 2001; Bugaj 2002; Bugaj 2004a) but, as yet, there has not been any attempt to thoroughly and systematically investigate the diversity of inflectional forms in OSc texts and investigate the factors conditioning its orthographic realisation.(1) lists six tokens of the plural noun *land* from various OSc legal manuscripts, taken from *A Linguistic Atlas of Older Scots* (LAOS), each with a distinct form of the {S} inflection, including a zero-morpheme, forms with covered inflectional <i>, <y> and <e>, a syncopated form with no covered inflectional vowel (CIV), and an abbreviated form <ſ>.

(1) <land>, <landſ>, <landes>, <landis>, <landys>, <landȝ>

In a manuscript note, Aitken (1977) stated that he had "regrettably not yet made the time to discuss [...] prefix and suffix syllables". Macafee, in her 2002 preface to Aitken's *The Older Scots Vowels*, elaborates that "without further data, [Aitken] did not feel that he could improve on the fullest account available, that of Kuipers (1964: 67-69)". Kuipers' account is a descriptive chapter within a larger work analysing two Eucharistic tracts written by Quintin Kennedy, a sixteenth-century Scottish abbot and religious reformist. Whilst Kuipers' treatment of the inflectional forms used in Kennedy's tracts is detailed and informative, its scope extends only as far as the work of the single scribe who is the subject of his study. Since the completion of LAOS, it has been possible to access more than 1000 legal texts in OSc as part of a lexico-grammatically tagged corpus. In this study, I present the LAOS data compiled by Williamson (2008) as precisely the "further data" which Aitken felt was lacking in 1977. Using data extracted from LAOS, I investigate the distribution of the various orthographic realisations of OSc {S} and {D}. The lexico-grammatical tagging of the LAOS data enables the near-instant identification of a large enough dataset of inflected tokens to perform detailed statistical analyses.

The results of these analyses cover the distribution of each type of inflectional realisation exemplified in (1), firstly considering the factors correlating with the use of zero {S} forms, then the abbreviation of {S} to <ſ>. Both {S} and {D} are investigated with regard to syncopated inflectional forms and the variation between the potential realisations of the CIV.

Lay Summary

The focus of my PhD research is the inflectional morphology of OSc. Specifically, I investigate the different ways scribes of fifteenth-century Scots legal documents wrote certain noun and verb inflections. (2) lists some of the spellings of the word *lands* which are found in OSc texts. There is a large amount of variation represented. For example, the vowel of the inflection is written as <e> in (2a), <i> in (2b), <y> in (2c) and omitted altogether in (2d). In (2e), the entire inflection is represented by the symbol <ſ>, and in (2f) it is not realised at all.

(2) a. <landes>

b. <landis>

c. <landys>

d. <landȝ>

e. <landſ>

f. <land>

Using LAOS, a corpus of OSc legal texts dating from 1380 to 1501 which have been tagged with linguistic information, I use statistical modelling techniques to analyse a large dataset of tokens from a a wide variety of legal texts dating from 1380 to 1501. The objective of this investigation is to draw conclusions as to why OSc scribes used particular orthographic variants of inflections.

Contents

Chapter 1

Introduction

1.1 Older Scots

Older Scots (OSc) is the label traditionally applied to the language spoken in Lowland Scotland in the period 1100 to 1700. First proposed by Aitken (1985: xiii), this period is broken down into further sub-periods:

(a) Pre-Literary Scots (1100-1375);

(b) Early Scots (1375-1450);

(c) Early Middle Scots (1450-1550); and

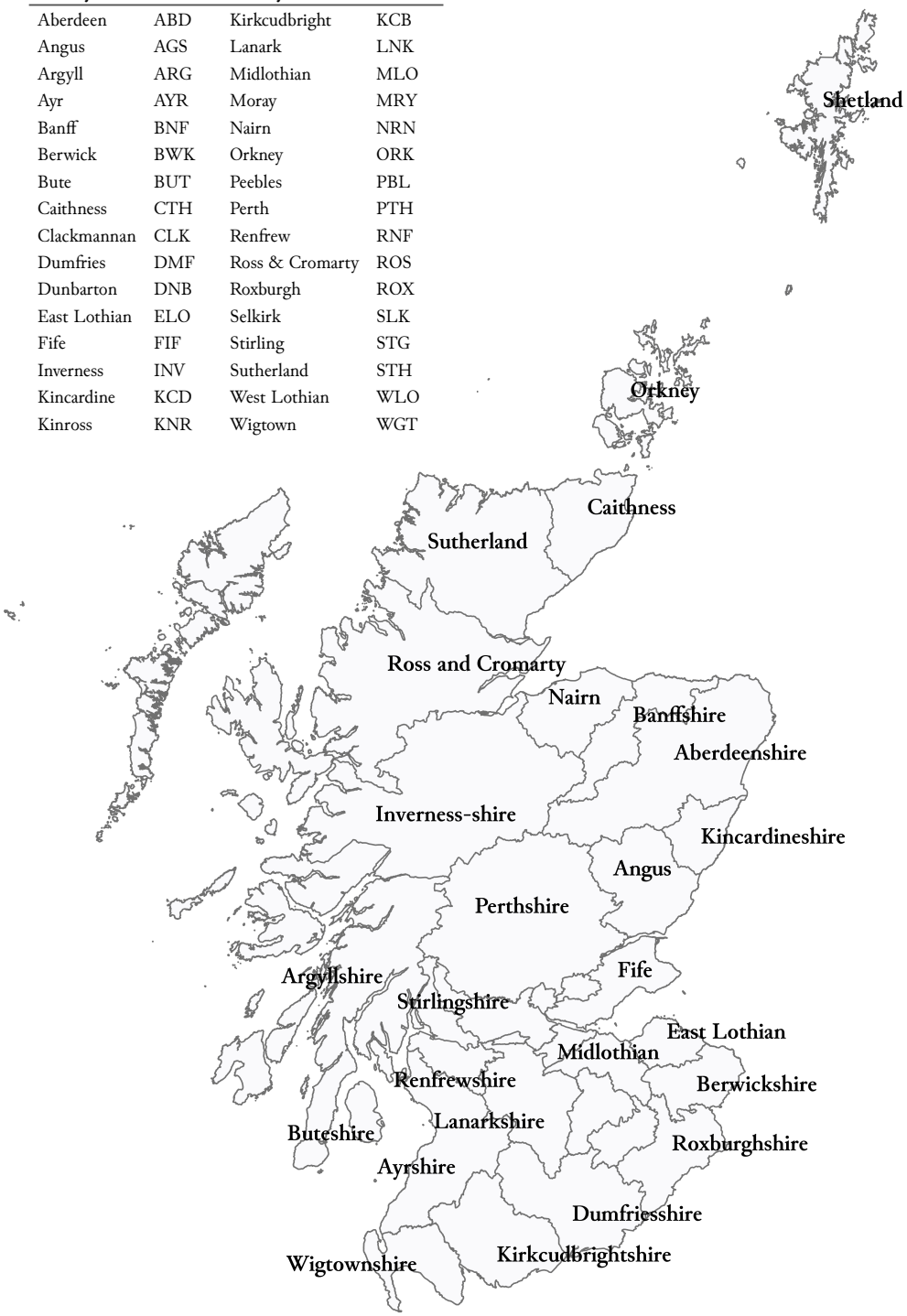
(d) Late Middle Scots (1550-1700).

This thesis is concerned with the period covered by *A Linguistic Atlas of Older Scots* (LAOS) (Williamson (2008)), 1380-1500 (see section 1.2). According to Aitken's model, this period falls between the Early and Early Middle periods of Scots, however, Williamson refers to the Scots exemplified by LAOS simply as 'Older Scots (OSc)'. I therefore adopt this label throughout this thesis, in which the term 'OSc' should be understood to refer to the period covered by LAOS.

The OSc period is roughly concurrent with the Middle English (ME) and Early Modern English (EModE) periods South of the Scottish-English border. OSc itself is derived from various dialects spoken by English settlers from around the twelfth century, including Anglian and Old Northumbrian. OSc is, however, a language distinct from these input sources, with various marked diagnostic features identified as characteristic of OSc as distinct from ME. Kniezsa (1997:41) identifies several features as diagnostic of the "Scottishness" of a text. One of these features is the orthographic realisation of the npl {S} morpheme (henceforth '{S}') as <is> or <ys>, as shown in (3).

FIGURE 1.1: The pre-1975 counties (*shires*) of Scotland, with a table showing the 3-letter abbreviation for each county used in LAOS.

County	Abbrev.	County	Abbrev.
Aberdeen	ABD	Kirkcudbright	KCB
Angus	AGS	Lanark	LNK
Argyll	ARG	Midlothian	MLO
Ayr	AYR	Moray	MRY
Banff	BNF	Nairn	NRN
Berwick	BWK	Orkney	ORK
Bute	BUT	Peebles	PBL
Caithness	CTH	Perth	PTH
Clackmannan	CLK	Renfrew	RNF
Dumfries	DMF	Ross & Cromarty	ROS
Dunbarton	DNB	Roxburgh	ROX
East Lothian	ELO	Selkirk	SLK
Fife	FIF	Stirling	STG
Inverness	INV	Sutherland	STH
Kincardine	KCD	West Lothian	WLO
Kinross	KNR	Wigtown	WGT



(3) <acct(i)onis causis & q(ua)rellis> (LAOS text 27 [Gordon Papers, 1491])
actions, causes and quarrels

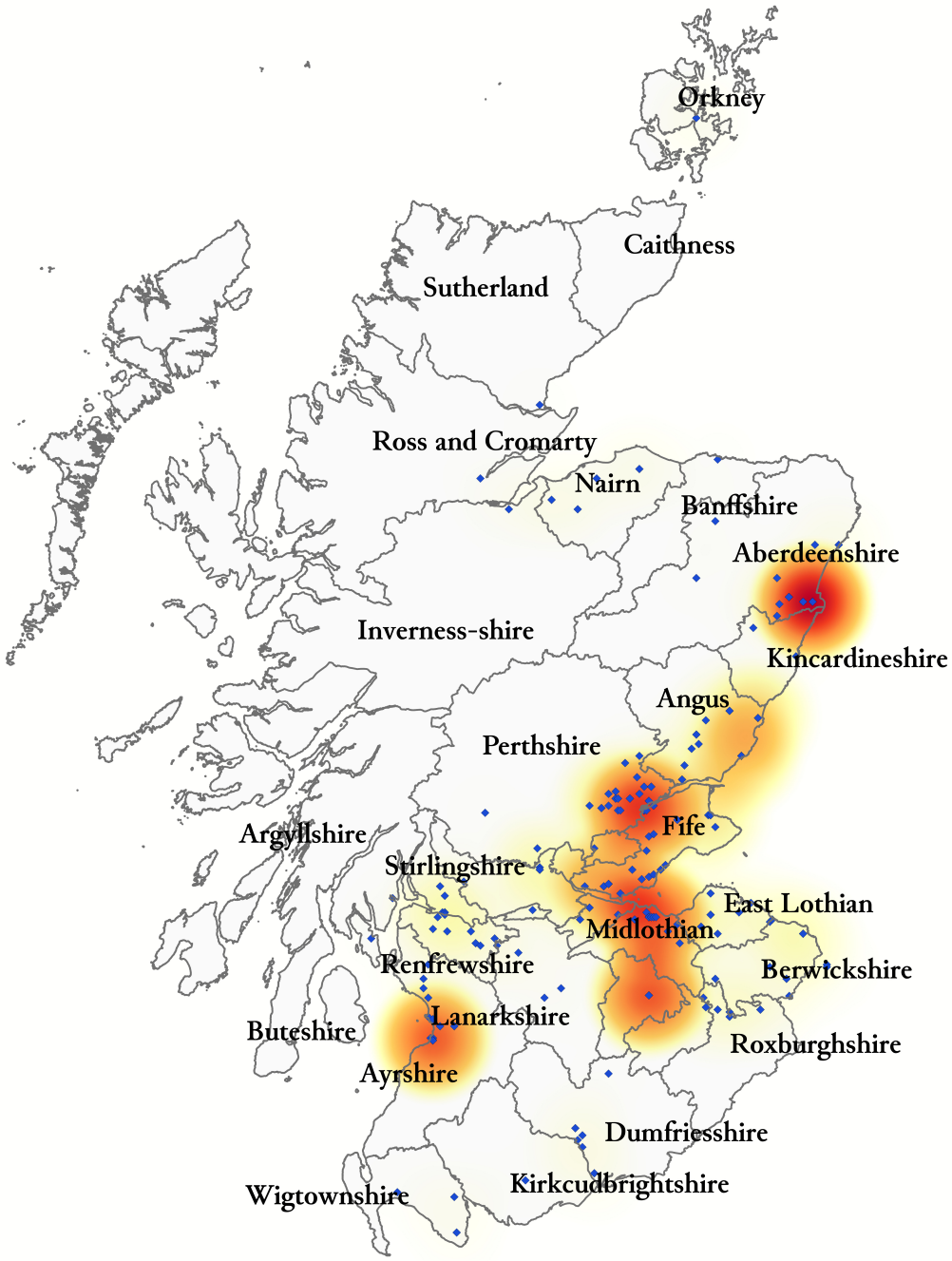
1.2 *A Linguistic Atlas of Older Scots* (LAOS)

LAOS (Williamson 2008) is an online, lexico-grammatically tagged corpus containing transcriptions of approximately 1,250 OSc legal records. These records are dated between 1380 and 1501, and represent various types of legal text, such as burgh records, cartularies and charters. The texts included in LAOS cover a wide geographical area, with the majority of texts localised to a specific county. The county names used in LAOS correspond to geographic divisions which were in official use in Scotland between 1890 and 1975. Figure 1.1 shows a map of Scotland with these boundaries indicated and labelled with the pre-1975 name of each county. Above the map is a table which lists the three-letter abbreviation for each county used in LAOS. These short county labels will be used throughout this thesis to indicate the geographic provenance of any LAOS texts mentioned. Figure 1.2 shows the same map again, this time with the location of each text included in LAOS represented by a single point. Where there are many texts localised to the same precise location (for example, in densely populated burghs such as Edinburgh), the points are overlaid on top of one another. A further 'heatmap' layer has been applied to indicate areas with the highest concentration of texts (the brightest red areas therefore have the most texts localised to them).

1.3 Statistical methods in historical dialectology

The recent availability of large-scale diachronic corpora has made it possible to investigate historical linguistic phenomena using sophisticated statistical methodology. Gries (2015: 97) describes corpus data in general as "observational and, thus, usually unbalanced and messy/noisy". He points out that techniques frequently employed in other areas of linguistics, particularly those which use experimental methods (such as sociolinguistics), can be applied to corpus linguistics. More than that, he suggests that these methods are actively necessitated by the very nature of corpus data. In particular, Gries suggests that mixed effects regression modelling is highly beneficial to corpus-based research due to the hierarchical nature of corpus data. For example, consider a corpus of 100 texts each containing 1,000 words. The corpus therefore contains a total of 100,000 words, each 'nested' within a particular text. An analysis using this corpus needs to account not only for the trends observable within the corpus as a whole, but for potential similarities between tokens drawn from the same text. Gries argues that this and similar sources of potential systematic variance (such as that between individual tokens of the same lexical item) are often overlooked in corpus studies, to the detriment of the conclusions drawn from them.

FIGURE 1.2: A map showing the texts included in LAOS. Individual texts are indicated by points, but as there are many texts localised to the exact same coordinates, a heatmap is also used to indicate text frequency.



Historical corpus studies are no exception to this generalisation, but historical dialectology is a field in which the potential for access to the kind of 'big data' necessary for the kind of analyses Gries advocates has been understandably smaller than in others. Compare, for example, the following corpora:

(a) **The Edinburgh Twitter Corpus (Petrović et al. 2010)**

A corpus of social media interactions containing 97 million tweets with a combined total of over 2 billion words.

(b) ***A Linguistic Atlas of Early Middle English*** **(LAEME)**

A corpus of ME texts containing approximately 300 transcribed medieval manuscripts with a combined total of approximately 650,000 words.

Petrović et al. (2010) corpus of tweets was extracted automatically over a period of two months using Twitter's streaming application programming interface (API). By contrast, LAEME was compiled over the course of 20 years by the manual sourcing, reading, deciphering and transcribing of manuscripts (Laing and Lass 2013b). That the Twitter corpus could be compiled in 1/120th of the time it took to compile LAEME, and in doing so contain around 150 times as many words, illustrates, albeit with an extreme example, the potential difference in data accessibility between historical linguistics and other fields such as sociolinguistics. Studies which require medieval manuscripts, or transcriptions thereof, as their primary source material are particularly susceptible to issues like this, due to the large amount of variability in spelling and orthographic conventions. Therefore, manual transcription is a necessity, rather than machine transcription.

The techniques which Gries (2015: 97) suggests are beneficial and even vital to corpus research have long been out of reach for historical dialectology. However, the completion of projects such as LAOS, LAEME, *An Electronic Version of A Linguistic Atlas of Late Medieval English* (eLALME) and *A Corpus of Narrative Etymologies* (CoNE) has rendered possible large-scale investigations of trends in medieval manuscript data using statistical methods.

1.4 Thesis structure

This investigation into the orthographic representation of OSc inflections is presented in three parts. Part I offers an overview of the existing scholarship on OSc inflections. To contextualise the topic, a broad overview is initially presented, beginning with the development of atonic vowels from Old English (OE) through to ME, and the processes of weakening and loss which eventually resulted in the inflectional systems of Modern Scots (ModSc) and Present Day English (PDE). The focus of the literature review is then narrowed to concentrate on the variety of ME which is geographically closest to OSc, Northern Middle English (NME).

The characteristically OSc covered inflectional <i> is introduced, along with its potential phonetic roots in NME pre-coronal raising. Once this background to the oft-cited OSc covered inflectional <i> has been introduced, the discussion moves onto more empirical ground, giving an account of the few studies which quantify the orthographic realisations of OSc {S} AND {D}. It emerges that there have been no large-scale empirical studies of the distribution of orthographic variants of OSc inflections. The arena in which to begin the current investigation, then, is something of a blank slate. Having said that, some particular areas of interest emerge from the received wisdom regarding OSc inflections together with the small amount of research which has been possible prior to the publication of LAOS. These are outlined in chapter 3. Part II firstly gives a detailed overview of the dataset which forms the basis of this investigation in chapter 4, and then in chapter 5 explains the statistical methodology which will be used. Part III presents the results of these analyses, organised into chapters according to the dependent variable (DV) of interest. Chapter 10 treats the insights gained about each DV in turn, reviewing the statistical results as well as expanding on the various focussed, qualitative analyses which these inspired.

1.5 The doctrine of *littera*

In the introduction to LAEME, Laing and Lass (2013a) address the difficulty of applying modern linguistic theoretical concepts such as *phoneme* and *grapheme* to medieval writing systems. They argue that such concepts imply a level of structuralism which is not representative of the medieval scribal tradition. In order to adequately characterise medieval scribal behaviours, Laing and Lass turn instead to a system they describe as "a theoretical framework and notation that cohere more closely with what scribes would have experienced in their education" - the doctrine of the *littera*.

Laing and Lass (2003) give a translation of what they describe as "the canonical definition [of *littera*] for medieval writers", that contained in the *Ars maior* of the fourth-century grammarian Aelius Donatus. The *Ars maior* was the second part of Donatus' *Ars grammatica*, a treatise on Latin grammar which was a fundamental part of the education of medieval scribes (Haugen 1950: 41; Laing and Lass 2003: 258; Archibald 2013: 186). According to Donatus' definition, the *littera* has three component parts: (a) name (*nomen*); (b) shape (*figura*); and (c) sound value (*potestas*).

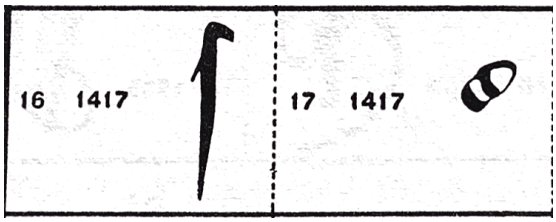
1.5.1 *Litteral* substitution sets (LSSs)

Litteral substitution in ME is defined by Lass et al. (2013) as follows: "Phonetic and orthographic remappings between Œand ME led to the availability, in a text community of multilingual scribes, of multiple orthographic variants to realise certain sounds. Conversely, multiple sounds could be mapped onto a single *littera* (see further Laing 1999, Laing and Lass 2003 and 2009)". That is, a particular *littera* could be used

by a scribe to represent several different sound values, and a particular sound value could be represented by several different *littera*. A group of *littera* which are 'interchangeable' in the sense that any member of the group can be used to represent any member of a corresponding group of sound values is defined by Laing and Lass (2003: 257-8) as an LSS.

In addition, a *littera* could have several *figural* representations. These are the different ways in which a scribe could form the shape of a given *littera*. Figure 1.3 shows an example of two variant forms of the <s> *littera* reproduced from Johnson and Jenkinson (1915: 44): the "long s", used word-initially and -medially; and "short s", used word-finally.

FIGURE 1.3: Illustration of variant orthographic forms of <s>, reproduced from Johnson and Jenkinson (1915: 44)



Throughout this investigation, I use the term *littera* to refer to single letters or abbreviation symbols as represented on a manuscript (MS) page regardless of the *figural* representation used by the scribe. In some cases, it becomes necessary to refer to the specific *figura* of a given *littera*, notably in my discussion of the palaeographic conditioning of the use of <ſ> as an abbreviation of {S} (section 7.2.1). Specifically, I refer to the potential forms of certain stem-final *littera* as conducive to particular inflectional realisations. I maintain the general use of *littera* throughout, however, in order to draw conclusions at the level of LAOS transcription conventions. The transcriptions in LAOS convey only the *littera* used by a scribe, not the detail of each *littera*'s *figural* realisation. Whilst I use facsimile examples from MS to illustrate various points relating to potential *figural* representations of *littera*, my conclusions on a wider scale remain confined to the *litteral* information contained in LAOS.

Part I

Previous Scholarship

Chapter 2

Literature Review

2.1 Introduction

In this chapter, I present an overview of the extant scholarship on the covered inflectional vowel (CIV) and, particularly, its representation in Older Scots (OSc) manuscripts (MSs). In section 2.2, I describe the reduction of the unstressed vowel system of Old English (OE) from four phonemes, /i/, /æ/, /u/ and /ɑ/ to the Middle English (ME) system of two-way contrast between vowel qualities represented orthographically as <e> and <i/y>, with the most frequently attested realisation of the CIV being <e>. I present the most dominant phonetic interpretation of this <e> *littera*, as a centralised schwa [ə]. This interpretation is then challenged on the grounds proposed by Lass (1992: 87). Lass's argument is that there is insufficient evidence to posit a centralised unstressed vowel at this point in the history of English, and that the manuscript evidence in fact suggests a merger in the unstressed vowel system to a more front vowel [e].

In section 2.2.3, I consider Lass' (1992: 78) suggestion that regionally specific orthographic CIV representations are indicative of "a range of allophonic variations" in the-realisation of the CIV. I focus particularly on the <i/y> variants of this vowel which appear in Northern Middle English (NME) texts alongside <e> forms. I then contrast interpretations of <i/y> as opposed to <e> in this position as orthographic variation with interpretations of it as an indication of a raised vowel quality.

In section 2.3, I place the previous discussion of NME <i/y> CIV forms in the context of OSc and the suggestion by King (1997: 161) that variable use of <i/y> and <e> as the CIV in NME indicates a parallel phonetic distinction. I argue that this interpretation posits phonetic variability for which the orthographic situation is insufficient evidence, pointing out the longer orthographic tradition existing in England leading up to the fifteenth century, as well as evidence from Lass (2009: 63) of a raised phonetic realisation of the unstressed vowel not represented by <i/y> in the orthography.

In section 2.4, I give an account of previous studies of the OSc CIV. I show that, whilst there is general agreement on the orthographic representation of {S} in OSc, its phonetic realisation is subject to debate. I also describe in this section a crucial difference in the realisation of the two grammatical categories of {S},plural noun (npl) and present tense verb (vps). Specfically, that vps inflections are subject to the Northern Subject Rule (NSR), a pattern whereby aspects of the verb's subject condition the realisation of {S}. I then go into more detail regarding the orthographic representation of {S} in OSc, presenting the conclusions drawn from empirical studies of Scots texts by Kopaczyk (2001; as Bugaj 2004a; 2004b). I address the general practice of expanding manuscript abbreviations to '-is' in editions of OSc texts, and suggest that the manuscript reality must be understood and documented before an account of the distribution of orthographic variations of OSc inflections can be produced. As well as this palaeographical issue, I address a methodological point arising from Kopaczyk's analyses, the status of an <e> *littera* where it can be interpreted as either stem-final or as part of the inflection.

Finally, in section 2.5, I discuss OSc {D}. Whilst issues of the phonetic quality and orthographic representation of the CIV have been largely covered in reference to {S} in the previous sections, {D} is notable in that it retains a raised realisation [ɪt] in some Modern Scots (ModSc) varieties (Macafee 1983).

2.2 Atonic vowels from Old to Middle English

2.2.1 Neutralisation

Minkova (2014: 227) states that the early OE system of unstressed vowels was, by the time of the first written evidence, composed of four phonemes: /i/, /æ/, /u/ and /ɑ/. Later in the OE period, these contrasts were reduced, evidenced by the "orthographic interchangeability" of vowel *litterae* in unstressed positions. Though vowel *litterae* were interchangeable, however, the most common representation of unstressed vowels was <e>, particularly in inflectional syllables Minkova (2014: 228). In derivational suffixes such as {ISH} < OE *-isc* (OE *Ænglisc*, ME *Englisch*, OSc *Inglis* Present Day English (PDE) *English*); and {ING} < OE *-ung, -ing* (OE *tidung*, ME *tidinge* PDE *tiding*(*s*)), the vowel of the suffix morpheme was <i> (Minkova 1991: 89), presumably indicating retention of a raised vowel quality.

The frequent representation of the CIV as <e> in ME has generally been interpreted as an indication of neutralisation in a mid-central schwa (Strang 1970: 290; Hogg 1992: 88; Minkova 2014: 228). Lass (1992: 78), however, points out that the use of <e> by Medieval scribes to represent the unstressed vowel surely suggests a merger in [e], or at least a merger in some vowel quality with a 'front' enough realisation to justify a large-scale generalisation of <e> where <a>, <u> and <o> were also available representations.

Lass et al. (2013) deliberately avoid the term "reduction" and the description of the unstressed vowel as

"central" in their definition of Weak Vowel Neutralisation[1]. They give two reasons for this decision. Firstly, they present evidence of the overwhelming popularity of <e> as the orthographic representation of the unstressed vowel in data extracted from *A Linguistic Atlas of Early Middle English* (LAEME) (Laing 2013). As an example, they state that 85% of strong past participle {EN} inflections in LAEME have <e> as the CIV. Secondly, Lass et al. point to contemporary descriptions of the phonetic realisation of the unstressed vowel, such as that of Hart (1569), which suggest that it was "qualitatively identical to short [e]".

2.2.2 Loss

Jordan and Crook (1974: 141) state that in Southern Middle English (SME), apocope of unstressed final vowels began in the twelfth century, taking place first in third syllables following long first syllables. In NME, however, the process began far earlier, with the loss of all final unstressed vowels complete by the thirteenth century. Minkova (2014: 230) suggests that the articulation of the final unstressed vowel was "an archaism" in NME by the mid-fourteenth century, though final <e> continues in the orthography far beyond this date. The relationship between final <e> and covered inflectional <e> is discussed in sections 2.4.3 and 4.1.3.

According to Jordan and Crook (1974: 141), CIVs in third syllables were the first to undergo syncope, beginning prior to the fourteenth century in NME, and later in SME. The npl {S} in an originally trisyllabic word such as *bishops* < OE *biscopas* would accordingly be reduced to [s/z] before the same morpheme in a disyllabic word such as *hounds* < OE *hundas*. Having said this, the same phonotactic constraints on this loss of syllabicity applied as in PDE, with the result that, where the loss of the CIV would lead to an illegal cluster, it was retained. Thus, {S} remains syllabic following [s, z, ʃ, ʒ, ʧ] and [ʤ], and {D} following [t] and [d].

Jordan and Crook do not speculate on the timeline of the loss of the CIV following its initial loss in trisyllables. Strang (1970: 180), however, suggests that, whilst CIV in disyllables was not lost immediately following the corresponding process in trisyllables, nevertheless there existed a "tendency to identify inflections wherever they occurred". This tendency led to the variable use of syllabic and non-syllabic forms of {S} and {D}, eventually giving way to the PDE system of retention of inflectional syllabicity only where the alternative is phonotactically disallowed.

2.2.2.1 Word class variation

Laing (2009), in a study based on LAEME, finds evidence for the conditioning of syncope in verbal inflections based on historical verb class membership and, most strongly, on a text's geographic location. In contrast to Jordan and Crook's (1974: 138) suggestion that inflectional syncope was in general more ad-

[1] Identified in *A Corpus of Narrative Etymologies* (CoNE), The Corpus of Changes (CC) as ((WVN)).

vanced in NME than in SME, Laing finds that NME texts show less evidence of specifically verbal syncope than SME texts. Having said this, the regional difference is most pronounced for vps forms, where the SME use of the {TH} suffix is particularly subject to syncope following dental stems. {TH} is far rarer in NME than it is in SME, with NME texts generally using {S} to indicate vps. Given that it is the npl {S} ending which Jordan and Crook specifically allude to in their account of ME affixal syncope, it may be that the comparative lack of syncope found in NME by Laing is representative only of differing vps marking strategies, which proceeded towards syncope at different rates. Where SME appears to have led the way in syncopating vps forms, it may be that the form used in SME conduced to syncope earlier than that used in NME. As well as this regional divide in the evidence for syncopated verb forms, Laing investigates the effect of stem-final consonants in conditioning syncope. As mentioned, dental stems are found to correlate with syncope of {TH}. Where syncope of {S} is found in NME texts, it correlates with stem-final [r] (Laing 2009: 261).

Though Laing does find some evidence of {D} syncope in NME, she states that it is rarer than in SME. The effect of stem-final consonants is also found to be significant. In particular, stem-final nasals and liquids are found to correlate strongly with {D} syncope. Minkova (2014: 232) also states that syncope of {D} was most common following [l] and [r], and least common following a consonant cluster. Though an analysis of the effect of stem-final clusters on verbal syncope is not presented by Laing (2009), she acknowledges consonant clusters as a likely conditioning factor for syncope, citing evidence that syllable weight can condition the occurrence of syncope.

In addition to these factors, Laing (2009: 261) finds a difference in frequency of syncope between past participle (vpp) and past tense verb (vpt) forms. Specifically, vpt forms are more likely to be syncopated. This suggestion is also made by Minkova (2014: 232), who suggests that it is due to the frequent adjectival uses of vpp forms, and points to the retention even in PDE of forms with non-syllabic vpt {D} and syllabic vpp {D} such as *aged*.

2.2.3 Orthographic variation

Lass (1992: 78) takes issue with the characterisation of the CIV in ME as "colourless" by Jordan and Crook (1974: 136). In addition to his suggestion that the most appropriate interpretation of covered inflectional <e> is something more like [e] than [ə] (see section 2.2.1), Lass points out that ME texts display regionally distributed traditions in the orthographic representation of the CIV which suggest "a range of allophonic distinctions". Specifically, Lass refers to the orthographic representation of this vowel as <u/o> in the South-West and West Midlands and as <i/y> in the North.

Lass contrasts two scholarly opinions regarding the orthographic representation of the CIV as <i/y> in NME texts: that <i/y> spellings indicate an attempt by scribes to represent a raised vowel quality [ɪ]

(Jordan and Crook 1974: 138; Strang 1970: 234; Minkova 1991: 121); and that the <i/y> spellings are simply orthographic variants of <e> and encode no phonetic difference. Lass et al. (2013) point out that ME orthography must be understood as one which is "not based on a phoneme-like 'transcriptional' praxis". That is, in a system where the particular quality and consequent realisation of the CIV does not in itself encode any morphological distinction, the realisation of the CIV as one particular character or another decreases in importance. One interpretation of this statement is that more than one orthographic realisation of the CIV may represent a single, neutralised vowel. As pointed out by Minkova (1991: 120), ME scribes using <i/y> as the CIV were employing the same vowel *littera* which they used to denote a high front vowel, including in derivational suffixes where the <i> spelling had been retained.

In Strang's (1970: 234) account, NME <i/y> CIV are explained as a neutralisation of the unstressed vowel system contrast between the [ə] of inflectional morphemes and the [ɪ] of derivational suffix morphemes, whereby all unstressed vowels preceding a final consonant were raised to [ɪ]. Whilst the suggestion that <i/y> in unstressed NME vowels represents a raised variant of an original mid vowel (whether that vowel is characterised as [ə] or [e]) is supported by others, including Lass (1992: 78), the reason given for the raised variant is not a neutralisation of contrast between raised and non-raised unstressed vowels. Rather, the environment in which the <i/y> variant appears (preceding <s>, <t> and <d>, presumably representing alveolar fricatives and plosives respectively) is identified as one conducive to vowel raising. As King (1997: 161) points out, the process of pre-dental raising in ME has been shown to extended to alveolar environments by Lass (1976: 185). Rather than previous [ə] being realised as [ɪ] in line with [ɪ] in other unstressed positions, <i/y> spellings in final closed syllables are an attempt by scribes to represent the product of Pre-Coronal Raising (PCR) in these environments (Minkova 1991: 120; Lass et al. 2013).

2.3 From Northern Middle English to Older Scots

This use of <i/y> in covered inflectional position in the North is acknowledged to be the source of the same variant's overwhelming use in OSc (King 1997; Kniezsa 1997). However, the phonetic implications of differing orthographic practice on opposite sides of the Scots border are debated. Kniezsa (1997: 41) lists <is/ys> inflections as diagnostic of a Medieval text's provenance as North or South of the Scots border. She does, however, point out that the use of <is/ys> variants in NME mean that this diagnostic of 'Scottishness' can often be contrasted only with SME counterparts. Having said this, King (1997: 161) suggests that a distinction can be drawn between OSc and NME texts based on the generalisation of <is/ys> in OSc as opposed to the variability between <is/ys> and <es> in NME.

King (1997: 161) compares the orthographic realisation of the CIV in NME with that in OSc, concluding that the <i/y> variant was used variably alongside <e> as the CIV in NME, whereas in OSc it dominated.

This contrast, King states, indicates a variability between [ə] and [ɪ] as the CIV in NME, and a generalisation of the [ɪ] resulting from PCR in OSc. Whilst this is a possible interpretation of the difference in orthographic representation of the CIV between NME and OSc, a factor which King does not consider is the longer history of English in Northern England before it spread to Scotland as the source of OSc. This longer history naturally meant a longer established orthographic tradition. In NME, the use of <e> in covered inflectional position, having a longer history and being more entrenched in the orthography, would not have been as easily displaced given an arguably small phonetic shift as it would be in Scotland, where vernacular writing was newer. In OSc, variable orthographic practice in the English input tradition would naturally generalise to the representation most aligned with the phonetic realisation. In NME, a sound change is clearly represented with <i/y>, but this new representation does not entirely replace the traditional <e>.

Furthermore, as Lass (2009: 63) points out, there are Middle English verse texts in which npl {S} written as <es> rhymes with *is*, the third-person singular vps form of *be*. This suggests that a raised variant of the unstressed vowel existed even where it was not represented in the orthography.

2.4 Older Scots {S}

In a manuscript note, Aitken (1977, quoted by Macafee 2002: xxvii) expressed regret that, despite his thorough study of the stressed vowels of OSc, he had "not yet made time to discuss [OSc] prefix and suffix syllables". As Macafee explains, "without further data, [Aitken] did not feel that he could improve on the fullest account available". The "fullest account" referred to by Aitken and Macafee is that contained in Kuipers' (1964) edition of two OSc eucharistic tracts composed in the first half of the sixteenth century. Kuipers states that, in the OSc texts he describes, the orthographic realisation of {S} is <is/ys>. There is widespread agreement that this form is typical of OSc. As mentioned in section 2.3, Kniezsa (1997: 41) describes it as a diagnostic feature of OSc. As well as its presumed graphic alternant <ys> (see section **??**, <is/ys> is also given as the main OSc realisation of {S} by King (1997: 160), Aitken and Macafee (2002: 71), Smith (2012: 45) and Bann and Corbett (2015: 5).

2.4.1 The phonetic realisation of OSc {S}

Npl {S} is described in the literature on OSc as generally being realised as <is/ys>, and infrequently as <es> (King 1997: 160; Aitken and Macafee 2002: 71), a description which agrees with that of Kuipers in reference to the tracts he describes. However, though scholars agree on the orthographic tendencies in OSc texts, this is not the case for the phonetic implications of these orthographic representations. In the Scots represented by the tracts he describes, Kuipers states, the phonetic realisation of {S} in OSc was much as it

is in ModSc. That is, [ɪz] after stem-final sibilants, elsewhere non-syllabic [s] after voiceless phonemes and [z] after voiced phonemes.

Aitken himself makes some "remarks on the history of *-is*" (Aitken and Macafee 2002: 71) based on verse evidence. His account agrees with Kuipers' in that npl {S} was syllabic after sibilants and non-syllabic after vowels, but as Kuipers' account is based on prose and Aitken's on verse, they do exhibit some differences. After non-sibilant consonants, Aitken states that there were certain environments where {S} was always non-syllabic, and others where syllabicity was optional. Where the inflection was syllabic, Aitken gives the vowel quality as /ɪ/. After an unstressed syllable, he states, the inflection was always non-syllabic, but after a stressed syllable it could be syllabic or not, dependent on the requirement of the poetic metre.

Aitken and Macafee's account, whilst it does not provide any empirical evidence of the variation in OSc inflections, is based on an extensive study of OSc poetry with a view to preparing it for performance (Caroline Macafee, personal communication, 27 May 2015). On the topic of optional syllabicity in inflectional syllables, Aitken and Macafee provide some general observations. They suggest that stem-final consonant realisation may have had an effect on the realisation of the inflection beyond the sibilant effect described by Kuipers, specifically that after a nasal or liquid consonant, the syllabicity of the inflection was sometimes preserved at the expense of a final unstressed stem syllable (for example, <eldris> *elders*). Aitken and Macafee do not claim that their generalisations are applicable to prose as well as verse. However, it is possible that stem-final phonology is conditioning factor for inflection realisation, as is shown to be the case for ME vpt and vpp by Laing (2009: 261) and OSc vpt and vpp by Macafee (1983) (see section 2.2.2.1).

In section 2.2.3, I presented an explanation of the NME <i/y> CIV which is the source of the same representation on OSc. This explanation pointed to a process of pre-coronal raising in covered inflectional position to account for the change from ME <e> to NME variable <e/i/y> and OSc <i/y>. Kuipers, however, asserts that, except in cases where the syllabicity of {S} is retained in PDE, the CIV had undergone syncope by the sixteenth century, a suggestion which is supported by Aitken and Macafee (2002: 71).

Whilst it is to be expected that the representation of the CIV should remain in the orthography of OSc for longer than it is realised phonetically (Bann and Corbett 2015: 5), the suggestion that the realisation of OSc {S} was identical to that of ModSc is questionable. Beal (1997: 341) quotes MacQueen (1957: 131) and Romaine (1982: 60), both of whom place the loss of orthographic <i/y> in the eighteenth century. If this is the case, Kuipers' claim requires the assumption that covered inflectional <i/y> was retained for almost two centuries after the vowel it represented had been lost from the pronunciation.

Having said this, the reduction and loss of the CIV which ultimately led to the ModSc realisation (or lack thereof) formed part of a gradual continuum of weakening and loss. The loss of the CIV in NME, as mentioned in section 2.2.2, began with syncope in trisyllables before 1300 (Jordan and Crook 1974) and syncope in disyllables was "underway by 1370" in NME (Strang 1970: 234). If the pattern of syncope

posited for NME were extended to OSc, then Kuipers' equation of the phonetic value of <is/ys> in the first half of the sixteenth century seems reasonable. Indeed, Strang goes on to suggest that npl and vps {S} had the "same shape" in late sixteenth century ME as in PDE. However, Murray (1873: 155), whilst he is in agreement with Kuipers that there was syncope of the CIV in early sixteenth century OSc, suggests that it was only in trisyllabic or longer words. Where {S} occurred with a monosyllabic stem, Murray suggests, it was independently syllabic.

In response to Murray's suggestions about the gradual syncope of the OSc CIV, Beal (1997: 341) offers an explanation of why this process might have proceeded at a slower rate in OSc than in NME. She suggests that the status of Scots as a standard language prior to the seventeenth century created a "constraint against the reduction of [ɪs]". The assumption, then, is that the continuum of weakening and loss of the CIV occurring in NME was not temporally paralleled by OSc, although the same outcome did eventually result.

2.4.2 Verbal and nominal {S}

Vps {S} is described in the literature on OSc as very similar in its realisation to the npl {S} inflection. Both are usually realised as <is/ys> and occasionally as <es>, the form more typical of their NME counterparts. The vps {S}, however, differs from ME not only in the representation of the CIV, but also in the representation of the inflectional consonant. In this respect, OSc differs only from SME. In SME, the consonant in the vps inflection is a dental fricative realised as <th/þ>. In NME and OSc, the consonant is <s>, as it is in npl {S}.

Where vps is represented using the {S} morpheme, it appears much like npl {S} in OSc. There is, however, a crucial difference between these two inflections in that vps {S} is subject to a rule conditioning its realisation dependent on subject type and adjacency.

2.4.2.1 The Northern Subject Rule

The NSR[2] describes a pattern in NME and OSc whereby the inflectional ending of first person singular and all plural vps forms is realised as zero or as <e> where the subject is:

(a) adjacent to the verb; and

(b) a personal pronoun (de Haas 2011: 175).

The below examples of NSR and non-NSR environments are taken from *A Linguistic Atlas of Older Scots* (LAOS). For a full introduction to LAOS, see chapter 4. For each example, the number following "LAOS" in parentheses represents the number assigned to the text in LAOS. Example (4) shows the vps inflection

[2]This phenomenon has also been referred to as the Northern Present Tense Rule and the Northern Personal Pronoun Rule (King 1997: 175).

of first-person singular *give*. (4a) shows the typical OSc {S} inflection realised as <ys>. (4b) shows the same verb, this time directly adjacent to the personal pronoun subject *I*, triggering the NSR and therefore a zero inflection. In contrast, the first-person singular verb *grant* in example (4b), which is not adjacent to a personal pronoun subject, retains the regular OSc {S} inflection in the form of the abbreviation symbol <ſ> (see section 3.1.3).

(4) First-person singular [text 253: 1466, CAR, FIF]:

a. <j will and be y(ir) p(rese)ntȝ graunt(is) and gevys full leiffe>
I will and by these presents grant and give full life…

b. j giff and graunt(is) for(e) me and my(n) ayr(is)
I give and grant for me and my heirs…

Example (5) shows the vps inflection of first-person plural *counsel*. The verb takes the regular OSc {S} inflection where it is adjacent to the npl subject *judges* in (5a) but has a zero ending when directly adjacent to the pronoun subject *we* in (5b).

(5) First-person plural [text 709: 1491, CHA, RNF]

a. <ve forsaid(is) jug(is) (con)sall(is) & ordanys>
we aforesaid judges counsel and ordain…

b. i(n) ty(m) to cu(m) we (con)sall & ordanis
in time to come we counsel and ordain…

Similarly, in example (6), The third-person plural verb *come* takes the regular OSc {S} inflection where it is adjacent to the npl subject *merchandise* in (6b) but has a zero ending when directly adjacent to the pronoun subject *they* in (6a).

(6) Third-person plural [text 9521: 1482, STA, MLO]:

a. <quhare ony victalis or m(er)chandise cu(m)mys>
where any victuals or merchandise come…

b. <franch me(n) quhe(n) yai cu(m) here>
French men when they come here…

2.4.3 Orthographic variation

There have been few empirical studies of OSc inflections, due to the lack of readily available corpora of OSc material prior to the recent publication of LAOS (Williamson 2008). However, there are two notable studies by Kopaczyk which make use of Scots border counties data contained in *A Linguistic Atlas of Late*

Medieval English (LALME) (Kopaczyk 2001) and an independent linguistic analysis of the *Wigtownshire Burgh Court Book* (as Bugaj 2004a).

Kopaczyk (2001) finds that all of the OSc texts use <is/-ys> as their primary variant. Half of the OSc texts use <es>, though not as a primary variant. In contrast, approximately half of the NME texts use <es> as their main variant. Kopaczyk's data support King's (1997: 160) assertion that OSc used the form <is/ys> for npl {S} in the majority of cases. Because the texts with <es> in Kopaczyk's data are both dated before 1400, they also support King's statement that <es> was used in the earliest OSc texts. Having said that, <es> does appear in the later OSc texts but is excluded from the final counts by Kopaczyk (2001: 136) due to it being a rare variant in those texts. Kopaczyk concludes from her analysis that OSc texts were more homogenous in their representation of the noun plural inflection, whereas NME texts were characterised by greater variety, particularly in their alternation between <es> and <is/ys>.

However, whilst Kopaczyk's conclusions are supported by her data, she acknowledges that there are only 4 OSc texts available in LALME for comparison against 43 NME texts. There is no objective reason to assume that these texts are representative of OSc as a whole, especially as they are specifically selected from the areas of Scotland closest to England, Kopaczyk's main aim being cross-border comparison. In fact, there is reason to suppose that the texts may not be representative of OSc inflectional orthography. Simpson (1973: 44) discusses the plural abbreviation symbol which he describes as "a looped vertical mark which usually indicates suspension of -is or -es" which was "especially common [...] in vernacular texts". This <ſ> abbreviation symbol has traditionally been subsumed under one or other of the fully-realised inflectional forms, hence Aitken and Macafee's (2002: 71) characterisation of it as "the manuscript abbreviation for -is". The <ſ> abbreviation is discussed by Kopaczyk (2001), who found that it was most common in Cumberland, but that the OSc MSs did not use it at all. Based on this finding, Kopaczyk speculates that the plural abbreviation symbol was not frequently used in Scotland and, on the evidence of the majority unabbreviated plural inflection form in Cumberland being <es>, that the abbreviation could be interpreted as standing for <es>. Having said this, Kopaczyk does acknowledge the impossibility of definitively assigning an equivalence relationship between the abbreviation and a particular orthographic representation of the inflection. Even if a scribe consistently used <es> whenever he did not abbreviate the inflection, it is not certain that he considered <ſ> a direct abbreviation of <es>. If the scribe also occasionally used <is/ys> forms, the uncertainty is even greater.

Kopaczyk (as Bugaj 2004a) performed a study of the npl {S} forms found in the *Wigtownshire Burgh Court Book*, finding that the most common noun plural inflection realisation was <is/ys> (59%), followed by <ſ> (40%), which is represented in the LALME Linguistic Profile (LP) for the *Wigtownshire Burgh Court Book* as italicised *-es*. She acknowledges that it is possible to argue for the interpretation of <ſ> as <is/ys> and a consequent change in the usual LALME practice of expanding this symbol as *-es*, but

ultimately comes to the same conclusion as in her previous (2001) study, that it is impossible to speculate on whether a scribe intended one or other fully realised inflection when he wrote the abbreviated form. As stated in the LALME documentation (McIntosh et al. 1986), the encoding of abbreviated forms of {S} as *-es* is intended to provide consistency across the corpus, rather than as a judgement indicating any particular underlying representation.

As well as pointing out the dangers of arbitrarily assigning equivalent full-inflection values to abbreviation symbols, Kopaczyk's studies expose another obstacle to analysis of OSc inflections. In the four texts from the border counties, she records that all used at least one <es> form, and three out of four also used at least one <e3> form. In contrast, only the two earliest texts use <s>, which is a rare variant in both. In the Wigtownshire manuscript, Kopaczyk states that only one <es> form and no <e3> forms are found. On this basis, Kopaczyk claims that the text shows no evidence of English scribal influence. However, the lack of <es> forms in the data is due to Kopaczyk's decision to analyse plural forms ending in <es> or <e3> as instances of <s> attached to otiose final <e>. For example, <partes> *parts* is analysed as the stem <parte> with the plural inflection <s>. It is worth noting, however, that the only examples of <s> attaching to <e>-final stems which Kopaczyk gives (<termes> *terms*, <soumes> *sums*, <partes> *parts* and <cwstowmes> *customs*) are loanwords derived from French. She does not specify whether this variation of {S} inflection is specific to French (or Romance in general) loans, or whether it also occurs with Germanic stems. This issue is further discussed in reference to the partitioning of morphemes in the LAOS data in section 4.1.3.

2.4.4 Manuscript abbreviation of {S}

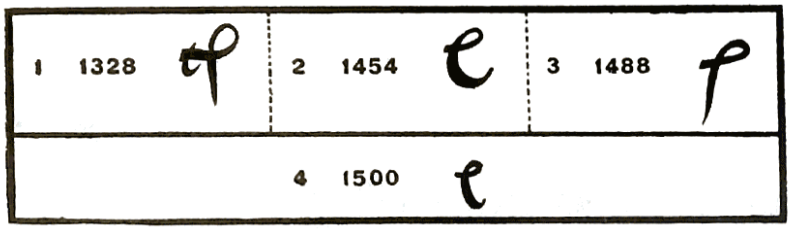
The tendency in the transcription of Medieval texts to treat manuscript abbreviations as irrelevant and to expand them silently is exemplified by Jenkinson's (1937) advice to transcribers of such texts. He advocates reproducing "all the peculiarities of the original" without making judgements as to their "value or interest". The single exception he makes to this rule, however, is to sanction the expansion of abbreviations "where there can be no doubt as to the way in which the original writer would have written them *in extenso*". In published editions of Medieval MSs intended to be read as prose texts, this procedure is likely to be both typographically necessary and reader-friendly. For the purpose of palaeographic and linguistic analyses, it masks the manuscript reality.

The three non-empirical accounts which directly address the OSc CIV (Kuipers 1964; King 1997; Aitken and Macafee 2002) examined in section 2.4 make no mention of manuscript abbreviation, aside from an acknowledgement by Aitken and Macafee of the existence of a "manuscript abbreviation for *-is*". Kopaczyk's empirical studies, on the other hand, present a mixed view of the role of abbreviation in the representation of {S} in OSc, with occurrence of abbreviation ranging from complete absence in the Scots-English border data (Kopaczyk 2001) through occasional attestation in the first part of the MSs of the *Wigtownshire Burgh*

Court Book, to accounting for 40% of all {S} inflections in the latter half of the same group of texts (as Bugaj 2004a; 2004b).

The abbreviation symbol itself is described in reference to ME usage by Johnson and Jenkinson (1915: 63). They classify the abbreviation as a "special sign" denoting "*-es* or *-is*, differentiating it from superscript letters and from mutually indistinguishable *sigla* used to indicate the suspension or contraction of a single *littera* or final groups of *litterae*. Their illustration of the form the abbreviation takes is reproduced in figure 2.1. Though the shape of the abbreviation appears fairly consistent, or, at least, recognisable, from the first illustration (1328) to the last (1500), Johnson and Jenkinson note its similarity to a stroke indicating suspension of final <g>, as well as other *litterae*. The most reliable criterion for identifying this abbreviation shape where it stands for *-is* or *-es*, they suggest, is whether the stem it occurs with is clearly a plural noun.

FIGURE 2.1: Illustration of the orthographic form of the {S} abbreviation, reproduced from Johnson and Jenkinson (1915: 63)



Simpson (1973: 44) describes the same abbreviation shape occurring in OSc texts indicating "suspension of *-is* or *-es*". Like Johnson and Jenkinson, Simpson suggests that the abbreviation was "especially common in plurals in vernacular texts". The roots of the abbreviation are, as suggested by Simpson's example of its use to signal the omission of the *is* in *multis*, in Medieval Latin orthography. Cappelli (1982: 2) describes it as a "truncation sign", stating that, though it could be used to signify the omission of any final letter or series of letters, it was most commonly used for *-is*.

2.5 Older Scots {D}

King (1997: 177) states that "the same considerations apply" to the CIV of OSc {D} as to that of {S}. That is, that the <i/y> graph represented [ɪ] when the inflection was syllabic. In her account of {S}, King (1997: 160) suggests that the phonetic realisation of the inflectional vowel occurred only in contexts where it occurs in ModSc, that is, in the avoidance of illegal clusters. The phonetic quality of the vowel is well-evidenced in this case, as the raised [ɪ] realisation of the vowel in {D} survives into ModSc, as shown by Macafee (1983). However, King's estimation of the progression of syncope in {D} is called into question by Macafee's assertion that varieties of ModSc retain [ɪ] following, not only [t] and [d], but all plosive

consonants.

This retention of the CIV in ModSc lends credence to the idea of a divergence of OSc from NME in terms of the CIV. In section 2.4.1, I referred to Beal's (1997: 341) suggestion that the status of OSc as a standard language prior to the seventeenth century may have inhibited the reduction and loss of [ɪ] in the npl and vps inflections. Together with the fact that, as shown by Laing (2009), more Northerly ME texts are less likely to show evidence of syncope in verbal inflections, the picture of OSc {D} is one of general resistance to {D} syncope in the varieties of ME closest to OSc and the codification of raised inflectional vowel forms in the OSc standard.

Macafee's (1983) account of the realisation of {D} as [ɪt] after plosives in varieties of ModSc is preceded by earlier accounts attesting the same phenomenon. Beal (1997: 351) compares the accounts of Murray (1873), Grant and Dixon (1921) and Dieth (1932), all of whom suggest that the transition from OSc to ModSc included the retention of [ɪt] after all plosive consonants.

Chapter 3

Research Questions

3.1 Methodological considerations

The comparatively recent availability of the *A Linguistic Atlas of Older Scots* (LAOS) corpus has made it possible to investigate, in a data-driven way, the realisation of Older Scots (OSc) inflectional morphemes. Extracting data from LAOS will be the first step in the investigation which I outline in part II. However, before embarking on an analysis of this data, it is necessary to address some methodological issues arising from my review of the literature in chapter 2. Whilst some suggestions and statements arising from the literature review in chapter 2 can straightforwardly be made the subject of investigation as independent variables (see section 3.2), there are three methodological points which stand out as obstacles to a thorough and consistent analysis of the data.

3.1.1 Ambiguous stem-final or covered inflectional <e>

The first of these methodological points is the status of the <e> grapheme in inflected forms with phonologically-consonant-final stems, such as <tymes> *times* and <juges> *judges*. The orthographic form of the inflection can, in these forms, be analysed as:

(a) <es>, attaching to consonant-final stems <tym> and <jug>; or

(b) <s>, attaching to <e>-final stems <tyme> and <juge>.

Both of these analyses are possible given what is known about OSc. Orthographic final <e> in uninflected verbs and nouns existed in OSc, though the corresponding phonological realisation is assumed to have disappeared by the time of the earliest OSc texts (see section 2.2.2). The evidence presented by Ackermann (1897: 15) suggests something in between these two representations. Ackermann suggests that after

unstressed syllables, final <e> had no meaning, and was used arbitrarily (*willkürlich*) as an orthographic addition to the end of a word. Where a word ended in a stressed, consonant-final syllable, however, Ackermann suggests that <e> was used as an indicator of vowel length and, consequently, was seldom used following stressed, consonant-final syllables with short vowel nuclei. More recently, Kniezsa (1997: 37) suggests that final "mute" <e> was used as a diacritic indicating the length of a stem-medial vowel. Kopaczyk (as Bugaj 2002: 87), however, suggests that final <e> in OSc manuscripts is more often an orthographic device than an indicator of vowel length.

As a first stage in my investigation, it will be necessary to decide on some consistent practice for separating stem from inflection in this kind of token. With this in mind, I analyse a relevant subset of the data in order to establish whether there are empirically justified grounds for one or other of the analyses described in item (a) and item (b).

3.1.2 Functional equivalence of covered inflectional <i> and <y>

The second methodological issue is the assumption that the graphemes <i> and <y> are interchangeable in unstressed position, with <i> being the dominant variant, giving way to <y> in contexts where <i> would result in an unbroken string of minim strokes such as stem-final <m>, <n> or <u> (Smith 2012: 29). Whilst this assumption is undeniably well-founded, it has not been empirically proven that the use of <y> as opposed to <i> in contexts where they are supposedly functionally equivalent is solely due to the occurrence of adjacent minim strokes. As part of the preparation for my analysis of the distribution of orthographic variants of {S} and {D} in LAOS, I perform an investigation into factors which correlate with the use of <y> over <i> as the covered inflectional vowel (CIV) to determine:

(a) whether the presence of adjacent <m>, <n> or <u> correlates with a higher likelihood of covered inflectional <y>; and

(b) whether there are other factors which correlate with a higher likelihood of covered inflectional <y>, such as geographical area or individual word.

3.1.3 Abbreviation of {S}

The final methodological issue is the status of the <ſ> *littera* used to represent {S} (see section **??**). It is described as an indicator of suspended final <is> or <es> (Johnson and Jenkinson 1915: 63; Simpson 1973: 44), though its original Latin usage, from which the vernacular use was adopted was, according to Cappelli (1982: 2), most commonly an indicator of <is> as a case marker.

The tendency in the literature on OSc inflectional morphology is to characterise <ſ> as a direct abbreviation of <is> and make no distinction between the two. My analysis will not follow this strategy, but

rather seek to describe and analyse accurate, *litteral* manuscript forms. Having said this, it is necessary to know whether the identification of <ſ> as a palaeographic device can be empirically verified. A preliminary analysis comparing fully-realised {S} forms with those realised as <ſ> will therefore determine:

(a) whether scribes employed <ſ> for {S} as a shorthand, convenient abbreviation of what would otherwise have been written fully as <is/ys>; or

(b) whether there is evidence in the LAOS data to suggest that factors other than scribal convenience motivated the use of <ſ>.

3.2 The distribution of orthographic forms

After these methodological points have been investigated and the data has been processed accordingly, I present an account of the distribution of orthographic forms of {S} and {D} which can be found in LAOS. I present these data firstly according to grammatical category in order to investigate whether the LAOS data reflect the characterisation of these inflections' orthographic forms in the literature. I then present the distribution of the same data with regard to several explanatory variables. Some of these variables are, as explained in the relevant section, motivated by my review of the literature. Others represent potential correlating factors which have not been previously considered. After the distribution of the data according to these variables has been outlined, I perform statistical analyses as outlined in Chapter **??** in order to ascertain which factors correlate with particular realisations of the CIV.

3.2.1 Geographic variation

Much of the literature on OSc inflectional forms focusses on the difference between OSc <i/y> CIV forms and Middle English (ME) <e> forms. A limitation of Kopaczyk's (as Bugaj 2004) investigation into the realisation of OSc plural noun (npl) {S} is the limited availability of OSc data (4 texts) in comparison to ME data (43 texts). As described in section 2.2.3, the aim of Kopaczyk's analysis is to investigate inflectional forms as part of a cross-border continuum, rather than to offer a full account of OSc inflectional morphology. As this investigation is able to take advantage of data drawn from the whole of Scotland, I show the geographic distribution of orthographic variants of {S} and {D}.

3.2.2 Temporal variation

King (1997: 160) suggests that, though <is/ys> was the most common realisation of {S} in OSc, "*-es* is the form we find in the earliest [OSc] texts". Though King: 160 does not explicitly say so, the implication is that <e> forms declined in OSc as the period went on. King also speculates on the phonetic loss of the

CIV in the OSc period. Though estimates of the date of complete orthographic loss of the CIV in OSc texts range from the sixteenth century (Aitken and Macafee 2002: 72) to the eighteenth (Beal 1997: 342), it may be that the LAOS data shows some early orthographic evidence of syncope.

I therefore present the temporal distribution of orthographic variants of {S} and {D}, showing:

(a) whether the LAOS data upholds King's claim that the earliest OSc texts have <e> as the CIV;

(b) whether, as King implies, these <e> forms give way swiftly to <i/y>; and

(c) if the data reveal forms of {S} and {D} with no orthographic CIV, whether the temporal distribution of these forms suggests that they may be evidence for the representation of syncope in the spoken language.

3.2.3 Lexical variation

Various aspects of an inflection's environment have been suggested to affect its realisation. Laing (2009: 261) suggests that the phonetic quality of certain stem-final consonants is conducive to syncope in ME verb inflections, even in Northern Middle English (NME) where, Laing states, syncope is generally resisted. In OSc, too, it has been suggested that stem-final consonant quality has an effect on the realisation of inflections. Beal (1997: 341) presents evidence for the retention of syllabicity in past tense and past participle {D} following stem-final plosive consonants long after the end of the OSc period, and Macafee (1983) confirms the existence of [ɪt] realisations of {D} in the same phonetic environment in dialects of Modern Scots (ModSc).

In the case of {S}, the trajectory of neutralisation and loss of the CIV is less clear. There is no evidence that an [ɪs]/[ɪz] realisation of {S} survives into ModSc, except, as in all varieties of English, where the inflection follows a sibilant consonant. Aitken and Macafee (2002) base their account of the phonetic realisation of {S} on verse evidence, a context in which the requirements of metre were likely to take precedence over the usual pronunciation of unstressed syllables. In verse, an inflection might be syllabic where it as not in speech or *vice versa*. However, Aitken and Macafee do point to a context where, if the omission of a syllable was required for metrical reasons, it was likely to be a final, unstressed stem syllable which underwent syncope as opposed to the inflectional syllable. Specifically, where the final unstressed syllable ended in a liquid or nasal consonant.

As well as stem-final consonant quality, the number of syllables in a word's stem may be a conditioning factor in inflection realisation. In ME, the loss of the covered inflectional vowel proceeded first in third syllables, that is, following disyllabic stems. According to Murray (1873: 155), this generalisation can also be applied to OSc, although he posits a later date for the eventual syncope of the CIV in monosyllables than is generally suggested for ME.

Part II

Methodology

Chapter 4

Corpus Data

4.1 *A Linguistic Atlas of Older Scots* (LAOS)

The compilation of *A Linguistic Atlas of Older Scots* (LAOS) (introduced in section 1.2) has allowed more detailed and wider-ranging study of Older Scots (OSc) orthography than was previously possible. Much of the study of OSc inflectional morphology has hitherto been based on verse evidence and, as demonstrated in chapter 2, assumptions about orthographic tendencies in OSc inflections have been used to justify phonological interpretations of the OSc covered inflectional vowel (CIV). By using the legal prose data provided in LAOS, I add to the picture of OSc inflectional orthography and consider what, if anything, the tendencies shown in legal documents add to our understanding of the phonology of the OSc CIV.

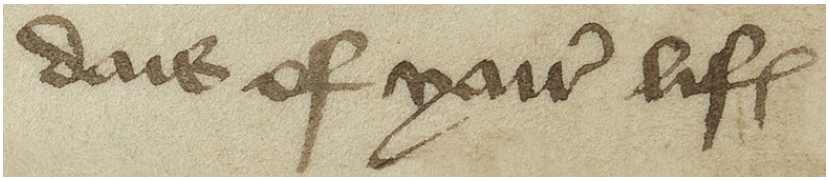
4.1.1 Transcription and tagging conventions

In this section, I firstly outline the transcription conventions employed in the LAOS corpus. By 'transcription conventions', I mean the system used by Williamson (2008) to encode the lexical, grammatical and orthographic information which makes up an individual LAOS token. Each token is a lexico-grammatically tagged representation of a single manuscript word form. After describing and exemplifying the general structure of LAOS tokens, I tabulate separately the conventions used to represent particular orthographic features of the manuscript text, and those used to encode grammatical information.

Each text in LAOS represents a diplomatic transcription of an OSc legal document. Each token of each word or—in some cases—morpheme is lexically and grammatically tagged. Figure 4.1 shows an extract from a text included in LAOS, with a transcription which aims to represent the manuscript characters, as well as a Modern Scots (ModSc) translation.

(7) shows the transcription of the phrase in LAOS, with each tagged token listed, as in LAOS, on a

FIGURE 4.1: Extract from LAOS text 36 [1447, CHA, ELO].



<dais of yaiꝛ liffꝭ> *days of their lives*

new line. Each token is prefixed with a dollar sign '$' and consists of: (a) a lexel, identifying the lexical item exemplified by the token; (b) a grammel, prefixed with a solidus '/', a code specifying the relevant grammatical information about the token; and (c) the transcribed form of the token as it appears in the manuscript, prefixed with a low line '_'.

(7) $day/npl_DAI+S $/pln_+S
$of/pr_OF
$/P23G_YAIR"
$life/npl_LIF+is $/pln_+is

The final word token in (7) has the orthographic form <liffꝭ>, as can be seen in the manuscript image in figure 4.1. This is represented, following LAOS transcription convention, as 'LIF+is', with LAOS upper case characters standing for manuscript lower case characters, and LAOS lower case characters standing for manuscript abbreviated characters. In the form 'LIF+is', the scribe has used a 'ꝭ' symbol to represent the noun plural {S} inflection. This has been transcribed in LAOS as lower case 'is', with the morphological boundary between the stem and inflection denoted by '+'. The separation of stems from affixes in this way is referred to as "hiving off" by Laing (2013) in the introduction to *A Linguistic Atlas of Early Middle English* (LAEME). Though this specific terminology is not employed by Williamson (2008), I use it here as a convenient term for the process of assigning a morpheme boundary within a transcribed word form. 'LIF+is' is a simple example with only a stem and a single inflectional suffix. (8) shows a more complex form with four separate morphemes including the stem, the plural adjective *unmovables*. In the LAOS transcription of this form, the boundaries between the morphemes {UN}, {MOVE}, {ABLE} and {S} are all indicated with '+', and the affixes are each tagged separately after the main tagged lexel.

(8) $move/ajpl<_VN+MOU+ABiL+IS $un-/xp-aj_VN+ $-able/xs-aj_+ABiL+IS $/plaj<_+IS

The methods used in LAOS to represent manuscript reality which are relevant for the current investigation are as follows:

(a) **upper case (UC) characters:** $lord/n_LORDE (figure 4.2a).

Upper case characters in LAOS represent fully-realised lower case manuscript characters. Some lower case manuscript characters are transcribed using lower case characters in LAOS. These are discussed in item (b).

(b) **Lower case characters**

1. conventional expansion of marks of suspension, contraction or truncation: $annual/aj_ANNUALE (figure 4.2b).
2. conventional expansion of systematic abbreviations such as <9> for {CON-} or {COM-}: $compear/vi_comPER (figure 4.2c).
3. a diacritic mark above a <y> indicated by 'x': $claim/npl_CLEM+YxS (figure 4.2d reproduced from Johnson and Jenkinson (1915: 54)).
4. some special fully-realised lower case characters:
   i. 'Ss' represents a double-s <ß>: $witness/n_WITNESs (figure 4.2e).
   ii. 'y' represents a fully-realised LC character, the shape of which is ambiguous between <y> and <þ>: $ordain/vpt_ORDAyN+D (figure 4.2f [reproduced from Johnson and Jenkinson (1915: 56)]).
   iii. 'z' represents either a yogh or a tailed 'z': <ȝ>. These characters are often orthographically near-identical. Their context, however, usually clarifies which letter is meant. In $year/npl_zER+is (figure 4.2g), the intention is clearly to represent an initial phonological semi-vowel in the word *years* using yogh. The same manuscript contains the token $letter/npl_LettRE+z (figure 4.2h), however, in this case, the grapheme represented in LAOS as 'z' is used to represent the sibilant segment of the noun plural {S} inflection.

(c) **Non-alphabetic characters**

1. two distinct shapes of scribal sigla are identified:
   i. tilde '~' represents a horizontal line above a word: $against/pr_AGAN~ (figure 4.2i).
   ii. a quotation mark represents an upward and backward curving line attached to a word-final letter: $thing/n_THING" (figure 4.2j).
2. morpheme boundaries are indicated using:
   i. a plus sign '+' for a contiguous morpheme boundary: $day/npl_DAY+IS (figure 4.2k).
   ii. a minus sign '-' for a non-contiguous morpheme boundary: $day/nG_DAY-IS (figure 4.2l).
3. superscript letters are indicated by a preceding caret '^'. For example, $ordain/vpp_ORDAN+^T (figure 4.2m).

4. characters judged by Williamson to represent insertions after the original word was written are enclosed between right-angle-brackets '<>': $call/vpp_CALL+>YT>.
5. Where one or more letters in a manuscript are difficult or impossible to decipher:
   i. characters which are completely indecipherable in the manuscript are indicated by an empty set of square brackets '[]': $commissar/npl_COmM[]SSAIR+is.
   ii. characters which are partially indecipherable are indicated by the probable characters enclosed in square brackets: $charge/vps_C[H]ARG+is.

The LAOS grammels which are used to tag the tokens relevant to this study are listed below. I have included only the 'base' form of the grammel, that is, the part of the grammel which specifies word class, number, case, person and tense. Many grammels in LAOS also include extensions specifying further grammatical information. These extensions will not be incorporated into this study, the reason for which is explained in section 4.1.2. The grammels included in the study represent:

(a) plural nouns: plural noun (npl)

(b) present-tense verbs: present tense verb (vps)
The base form 'vps' is followed by two numbers indicating:

1. sg. or pl. number (1, 2)
2. person (1, 2, 3)

For example, vps13 indicates a third-person singular form and vps21 a first-person plural form. An exception to this is vps tokens with 0 rather than 1 or 2 as their grammatical number specifier. This 0 indicates the formal use of the plural with a singular subject, as exemplified by (9).

(9) <we [...] be y(ir) oure l(ett)rez **Remittis**> [text 740: 1494, STA, INV]
we [King James IV] [...] *by these our letters* ***remits***

(c) past tense verbs: past tense verb (vpt)

(d) past participles: past participle (vpp)

The representations I have chosen to use for the corresponding morpheme labels are {S} for npl and vps inflections, and {D} for vpt and vpp inflections.

4.1.2 Data identification and extraction

The relevant data were extracted from LAOS using an automatic search for all tokens tagged with grammels beginning with any of the strings 'npl', 'vps', 'vpp' or 'vpt'. This search yielded all tokens tagged with these grammels as well as any extensions. In this section, I describe the changes I have made to the raw dataset and the reasons for these changes. Firstly, I list the types of token which were discarded at the outset[1]. Secondly, I explain the changes I have made to the transcription conventions of LAOS and why these changes are necessary to the current investigation. For ease of reference, I refer to the subset of LAOS data which forms the basis of this investigation as *Inflections in A Linguistic Atlas of Older Scots* (INFLAOS).

The following is a list of reasons for discarding subsets of tokens during the compilation of INFLAOS:

(a) **Superfluous grammels**

The initial-character search described in section 4.1.2 also yielded tokens with the grammel vpsp, indicating present participle. As present participle inflections do not form part of the current study, these tokens were manually removed from the dataset. In section 4.1.2, I mentioned that many grammels have extensions beyond the base forms listed, and that these extensions are not to be utilised in this study. In the majority of cases, these extended grammels can be reduced to their base forms simply by omitting the extension. For example, $croft/npl{o} includes the extension {o}, which indicates an onomastic use of the plural noun *crofts*. This grammel is included in INFLAOS simply as $croft/npl. However, some vpp grammels include the extension -aj or -av, indicating that the token has the form of a past participle verb but is in fact acting as an adjective (-aj) or adverb (-av). (10) shows an example of the lexel *expreme* used in a vpp-aj context, with (11) showing the same lexel in a vpp context in the same text for comparison. Vpp-av only appears twice in LAOS, as two tokens of *divisibly*, both part of the phrase *jointly and divisibly*, which seems to be similar in meaning to the ModSc legal term 'jointly and severally'.

(10) <ye la(n)d<ſ> befor expremit in man(er) & formꝫ>
the lands before expressed in manner and form [text 746: 1482, CHA, MRY]

(11) <ye said thanė his lachfull opynioñ or frend<ſ> as in the first band is expremit>
the said thane his lawful opinion as in the first bond is expressed [text 746: 1482, CHA, MRY]

(b) **Vowel-final stems**

Because the purpose of this investigation is to draw conclusions about the orthographic forms of inflections, and to use those conclusions to hypothesise about their phonetic realisation, it will be

[1] Other tokens were omitted at a later stage in the process of constructing the dependent and independent variables for investigation. These tokens are discussed in sections 4.1.4 and 4.1.5 and enumerated in Appendix A.

crucial to distinguish these morphemes from the stems to which they are attached. (7) and (8) above demonstrate how this 'hiving-off' is represented in LAOS. However, where a stem would be expected to be realised as phonetically vowel-final, this creates problems. Consider the tokens in (12) and (13).

(12) $assignee/npl

a. ASSIGNA+S

b. ASSIGNA+IS

c. ASSYNGAY+S

d. ASSIGNA+YS

In (12a) and (12b), the morpheme boundary is transcribed after stem-final <a>, giving the inflectional forms <s> and <is> respectively. In (12a), this is the only reasonable boundary because, given that <-as> is never attested in LAOS as an orthographic realisation of {S}, it is highly unlikely that the <a> belongs to the inflection <as> rather than to the stem as a representation of a phonetic final vowel. The placement of the boundary in (12b), however, is a matter of judgement. Singular *assignee* is attested in LAOS 35 times, including eight tokens with final <a> and six with final <ai>. Based on attested forms of the singular, then, there is reason to suggest that the stem form in the plural could be <assigna> or <assignai>. LAOS also contains the plural form <assignaijs>, with the morpheme boundary placed to indicate stem-final <ai> followed by plural <js>. In (12c) and (12d), the underlying manuscript form is identical, but the hiving off of the inflection in (12c) assumes stem-final <a> whereas that in (12c) assumes stem-final <ay>, both of which are attested as final strings in singular *assignee*.

(13) $pay/vpp

a. PA+YT

b. PAY+T

c. PAY+EDE

d. PAID

The appropriate placement of morpheme boundaries in such orthographic contexts is not always intuitively clear. This is attested to by the small number of instances in LAOS where identical manuscript forms have their inflections hived off differently, as in (13a) and (13b), where two tokens realised as <payt> are hived off differently, one with a stem-final <a> and the other with a stem-final <ay>. (13c) is hived off to indicate a stem-final <ay> followed by an inflection <ede>. However, it is impossible to rule out stem-final <aye> given the attestation of uninflected *pay* realised as <paye>.

(13d) has not been hived at all, possibly suggesting an interpretation closer to PDE wherein the boundary between the stem-final vowel and the inflection is not preserved.

Any decision about how to consistently deal with phonetically vowel-final stems such as these would necessitate making arbitrary assumptions about OSc scribes' practices in representing morpheme boundaries after vocalic stems. Instead, I made the decision to omit tokens of lexels which I judged to be phonetically vowel-final in OSc. (see Appendix A). In some cases, it was not clear whether a lexel should be categorised as vowel- or consonant-final. Items ending in <w> such as *know* and *allow* are phonetically vowel-final in ModSc, being realised with final [aʊ]. It is possible that in some syllable-final environments, OSc retained a semivowel [w]. However, Johnston (1997: 110) suggests that if this was the case, it was confined to a small geographic area and did not persist past the middle of the OSc period. I have therefore omitted such tokens under the assumption that they are phonetically vowel-final (see Appendix A). There are also cases where a vowel-final verb is used to lexically tag a consonant-final derived nominal form such as *do ~ doer*. These tokens are dealt with in section 4.1.5.

(c) **Irregular lexels**

This study aims to investigate the realisation of the regular npl, vps and vpt inflections, and the weak vpp inflection. Therefore, nouns and verbs which have irregular forms of these grammatical categories should be excluded. Table 4.1 lists irregular noun and verb lexels omitted from INFLAOS. Most of the nouns in table 4.1 form irregular plurals in OSc in the same way as they do in ModSc, either with a medial vowel change (sg. *man*, pl. *men*), an irregular plural suffix (sg. *ox*, pl. *oxen*) or no form change (sg. *fish*, pl. *fish*), but two have been regularised in ModSc: *cow* and *shoe*. The OSc plural forms of these are *ky* and *shoen* respectively. Of the verbal forms listed in table 4.1, most are irregular only in their vpt and vpp forms (for example, $become/vps indicates regular *becomes* whereas $become/vpt and $become/vpp indicate irregular *became*). Only *be* and *have* are irregular in their vps forms as well. Of the verbs with irregular vpt and vpp forms, most are equivalent in their irregularity to their ModSc counterparts. The only notable cases are *owe* and *work*, the vpt and vpp forms of which refer to *ought* and *wrought* rather than *owed* and *worked* respectively.

(d) **Unsuitable texts**

There are three texts in the corpus which I judged unsuitable to include in my analysis: one on the grounds of its date and type, and two on the grounds of their place of origin. The first is text 160, *The Scone Gloss*, which I have excluded on the grounds that it is not only typologically distinct from all other texts in the corpus ("Scots words and phrases written as interlinear glosses to Latin words in a charter" (Williamson 2008)), but it is also dated 20 years earlier than the next earliest text in the

TABLE 4.1: Irregular noun and verb lexels omitted from the dataset. Where a prefix is enclosed in square brackets, both the prefixed and the base form are attested. [-]man refers to the lexel *man* as well as compounds with *-man* as the head: *burlawman*, *countryman*, *craftsman*, *englishman*, *erseman*, *frachtman*, *frauchtman*, *freeman*, *gentleman*, *goodman*, *herdsman*, *husbandman*, *inbornman*, *kinsman*, *kirkman*, *leperman*, *marchman*, *overman*, *scotsman*, *shipman*, *sokeman*, *watchman*, *witnessman* and *workman*. [-]woman refers to the lexels *woman* and *gentlewoman*.

npl	**vps**	**vpt/vpp**				
child	be	[[um-]be-]think	begin	gar	mett	shoot
cow	have	[a-]bide	bide	get	misset	show
deer		[be-]come	bind	grant	oblige	sit
fish		[be-]seek	break	grow	ordain	slay
foot		[for-]bear	bring	gutter	owe	solicit
goose		[for-]bid	call	hang	poind	speak
kye		[for-]give	cast	have	present	steal
[-]man		[for-]go	charge	hear	prove	streek
ox		[out-]quit	clad	hecht	raise	strike
sheep		[out-]red	compear	heicht	read	subscribe
shoe		[out-]redd	contain	help	receive	summon
swine		[out-]run	cost	hold	rede	swear
[-]woman		[out-]take	deal	hurt	repute	threat
		[to-]put	dispone	inquiet	resign	vest
		[under-]stand	do	invest	ride	wet
		[wad-]/[re-]set	draw	join	rin	win
		act	enter	know	rise	wit
		append	escheat	lead	scathe	work
		appoint	fall	leave	scot	write
		assyth	feft	lend	see	wrought
		attend	fight	let	sell	
		be	find	lot	send	
		become	foresaid	make	shear	

corpus. Text 8001 is localised by Williamson to south of the Scottish-English border (Durham) and is therefore omitted from my analysis (see Appendix A). Similarly, text 88 contains not only the OSc endorsement for which it is included in LAOS, but also a portion of text, "the language of which is markedly different and is clearly English rather than Scots" (Index of Sources (IoS)). I have therefore omitted all tokens which come from the Middle English (ME) part of the document (see Appendix A), but have retained the OSc tokens under the same text number.

4.1.2.1 Changes to the LAOS transcription conventions

LAOS was compiled with a view to making the extant corpus of OSc legal records available to researchers investigating these texts from all linguistic angles. The meticulous transcription undertaken by Williamson (2008) succeeds in providing a transparent representation of the manuscript reality, whilst at the same time rendering possible detailed linguistic study. Of course, a resource of this magnitude cannot be suited in every way to every use which researchers wish to make of it. In this study, I use it as a resource to

investigate a specific phenomenon: the orthographic representation of certain inflections. Moreover, I apply to INFLAOS a particular type of quantitative methodology which has hitherto seldom been applied to historical corpus research. Because the transcription of the original documents was not completed with the requirements of this methodology in mind, cases arise where the transcription or encoding of LAOS tokens is incompatible in some way with the methodology. As such, it has in some cases been necessary for me to make use of different conventions to those used by Williamson.

A case in point is Williamson's use of lower case characters to represent manuscript abbreviations. This convention means that, for example, a token including a noun plural {S} inflection, fully-realised as an <i> graph followed by an <s> graph (figure 4.3a), is represented in LAOS as 'IS', whereas the abbreviation symbol <ſ> (figure 4.3b) is represented in LAOS as 'is'. Whilst this convention is perfectly adequate and clear when perusing a transcribed text itself, and also for investigations not concerned with the specific marks on the page, parts of this study rely heavily on an accurate differentiation of abbreviated and non-abbreviated inflectional forms. Any attempt to analyse or manipulate this data using a system which does not recognise or preserve the distinction between 'IS' and 'is' will therefore immediately run into problems. Table 4.2 gives an account of all changes I have made to the format of the data extracted from LAOS. For the remainder of this work, I refer to manuscript forms using these altered conventions, including where they appear as part of a tagged token. For example, a token represented in LAOS as $cunzer/npl_CUnzEOURis will be referred to henceforth as <cu(n)ȝeourſ> *cunzers*.

4.1.3 Ambiguous <e>

The variable use of final <e> in singular nouns in OSc (see section 2.2) creates ambiguity about the linguistic status of <e> in words like *gudes*. Consider the forms of $time/npl in example (14), which are taken from LAOS and are presented here without their morpheme boundaries as they are placed in LAOS omitted.

(14) a. <tymis>
b. <tyms>
c. <tymees>
d. <tymes>

In (14a) and (14b), {S} appears to be represented by <is> and <s> respectively. In example (14c), the inflection is presumably <es>, attached to a form with stem-final <e>. In example (14d) however, it is not clear whether the inflection should be interpreted as <s> attached to an e-final stem (<tyme#s>), or <es> attached to a consonant-final stem (<tym#es>). As a singular noun, *time* is attested in LAOS with final <e> in approximately half of all tokens, and is therefore unenlightening on the subject of where to place the boundary between the stem and {S}. LAOS itself is occasionally inconsistent in this regard, with some

TABLE 4.2: List of changes to LAOS transcription conventions.

LAOS	**MS reality**	**Current study**
UC characters	LC characters	LC characters
LC characters	abbreviation symbols, marks of suspension, contraction or truncation	LC characters in parentheses
Ss	double-s <ß>	ß
z	tailed z <ʒ> or yogh <ȝ>	ȝ
tilde ~	horizontal line above word form	See section 4.1.4.
quotation mark ”	upward and backward curving line attached to word-final grapheme	See section 4.1.4.
+	contiguous morpheme boundary	#
-	non-contiguous morpheme boundary	There are very few tokens with non-contiguous inflection boundaries in INFLAOS. These two categories are therefore merged under the representation #.
>x> or [x]	insertions and difficult to decipher characters	INFLAOS contains only one instance of >inserted text> which is omitted. Where characters are marked as [difficult to decipher], I have trusted the judgement of Williamson and conflated these with the corresponding unbracketed characters.

identical tokens hived off to indicate a stem-final <e> (<gude#s> *goods*) and others to indicate a covered inflectional <e> (<gud#es>). In order to have a consistent system of representation for forms which have this ambiguous <e>, I have changed all examples like <tyme#s> and <gude#s> to <tym#es> and <gud#es> etc. As a result, there are now no tokens in which a potentially stem-final <e> is not treated as part of the inflection. There are, however, tokens with stem-final <e> which are followed by another vowel, such as example (14c).

(a) identify any tokens hived off in LAOS in such a way that the *littera* preceding the morpheme boundary is <e> and that following the morpheme boundary is an inflectional consonant. For example, <gude#s> *goods*.

(b) remove the morpheme boundary and replace it preceding the <e>. For example, <gud#es> *goods*.

After this procedure is completed for the whole dataset, there are no remaining tokens in which an ambiguous stem-final or inflectional <e> is not hived off as part of the inflection. There are, however, tokens with stem-final <e> which are followed by another vowel, such as example (14c). These are discussed further in section 4.1.2.

4.1.4 Dependent variables

To successfully investigate the distribution of orthographic variants of the npl, vps, vpp and vpt suffixes, the dependent variables (DVs) must clearly represent and distinguish these orthographic variants. In total, LAOS contains 39 unique forms of {S} and 43 unique forms of {D}.. Table 4.3 represents the structure of LAOS npl and vps {S} tokens as consonant and vowel *littera* strings. Table 4.4 shows the structure of OSc vpt and vpp {D} in the same way.

TABLE 4.3: The structure of OSc {S} tokens deconstructed into consonant and vowel *littera* strings. A baseline asterisk '∗' signifies a diacritic mark or siglum. 'Form' lists the most common (or only) form of the string attested in LAOS.

String	Form	npl tokens	vps tokens
ſ	<ſ>	7,728 (53%)	1,017 (34%)
VC	<is>	4,913 (34%)	1,007 (34%)
(zero)		1,361 (9%)	915 (31%)
C	<s>	327 (2%)	21 (1%)
C∗	<s->	171 (1%)	6 (0%)
VC∗	<eʒ->	22 (0%)	0 (0%)
V∗C	<yxs>	7 (0%)	4 (0%)
VCC	<ysß>	4 (0%)	1 (0%)
VVC	<iis>	3 (0%)	0 (0%)
CC	<sß>	1 (0%)	0 (0%)
VCe	<ese>	1 (0%)	0 (0%)
V∗	<yx>	1 (0%)	0 (0%)
Total		**14,539**	**2,971**

TABLE 4.4: The structure of OSc {D} tokens deconstructed into consonant and vowel *litterae* strings. A baseline asterisk '∗' signifies a diacritic mark or siglum. 'Form' lists the most common (or only) form of the string attested in LAOS

String	Form	vpp tokens	vpt tokens
VC	<it>	3,919 (83%)	1,446 (84%)
C	<t>	671 (14%)	245 (14%)
Ce	<de>	76 (2%)	8 (0%)
VCe	<yte>	22 (0%)	1 (0%)
VC∗	<it">	12 (0%)	3 (0%)
V∗C	<yxt>	9 (0%)	0 (0%)
C∗	<t">	7 (0%)	3 (0%)
VCC	<itt>	3 (0%)	2 (0%)
VVC	<zyt>	3 (0%)	3 (0%)
hVC	<hit>	2 (0%)	0 (0%)
CC	<th>	0 (0%)	1 (0%)
VVC∗	<zyt">	0 (0%)	1 (0%)
Total		**4,724**	**1,713**

The attested orthographic realisations in LAOS of {S} inflections can therefore be represented as:

{S} → abbreviation symbol / zero / (V(V/∗))(C(C/e/∗))

That is, {S} can be realised as: (a) an abbreviation symbol; (b) a zero morpheme; or (c) a string of *litterae*, generally consisting of a vowel and a consonant or a consonant only, and potentially also an additional covered vowel, an additional consonant, a final <e> or a siglum (represented here with a baseline asterisk '∗').

Table 4.3 shows that the majority of npl {S} tokens are realised as an abbreviation symbol (53%). Of the 8,745 tokens of {S} realised as an abbreviation symbol, 8,744 are realised as <ſ> and one as <(us)>. Figure 4.4 reproduces an example given by Cappelli (1982: 13) of the realisation of the <(us)> abbreviation in Medieval Latin palaeography. Cappelli states that this same <9> symbol could also be used as an abbreviation of <is> and <s>. For the purpose of this study, the single <(us)> token is discarded and the abbreviation symbol represented henceforth as <ſ>.

Both npl and vps {S} is also often represented by <VC> (34%). The high percentage of vps tokens with zero inflection is due to the operation of the Northern Subject Rule (NSR) (see section 2.4.2.1, and further discussion in chapter 6). The realisation of {S} includes <CC> in only five tokens, and <VV> in only three. As the low frequency of these tokens disallows analysis of whether they behave in a statistically different way to single-V or single-C inflections, they are omitted from the dataset (see Appendix A). The same is true of the single <V∗> token and the single <VCe> token. Tokens which contain a siglum or diacritic (signified in the formula above by a baseline asterisk '∗') are more numerous, but as they share a basic form with their 'no-siglum' counterparts, they are merged into the same category and the presence of a siglum recorded as a separate variable for the token. An example of this reclassification of inflections with final *sigla* is shown in table 4.5. Table 4.5a shows the original classification of inflectional VC strings for three tokens of npl *article* and three tokens of vps *happen*. The last column in table 4.5b shows the new variable inflection-final *siglum* (IFS) which specifies the type of *siglum* which follows the inflection. The inclusion of this separate variable preserves the *siglum* information whilst allowing inflections with the same *litteral* string to be combined in a single category.

The attested realisations of {D} inflections can be represented as:

{D} → zero / (h)(V(V/∗))(C(C/e/∗))

That is, {D} can be realised as: (a) a zero morpheme; or (b) a string of *litterae*, generally consisting of a vowel and a consonant or a consonant only, and potentially also an initial <h>, an additional covered vowel, an additional consonant, a final <e> or a siglum.

Table 4.4 shows that the majority of vpt and vpp {D} is realised as <VC> (83-84%). Those tokens which are not realised as <VC> are most often realised as <C>, with fewer than 3% of tokens realised as anything

TABLE 4.5: The original classification of inflectional VC strings for three tokens of npl *article* and three tokens of vps *happen*, and the resulting classification after encoding *sigla* as a separate variable.

(A) Before

Text	Lexel	Grammel	Form	Inflection	String
T752	*article*	npl	articlis	is	VC
T496			articliīs	iīs	VC∗
T1316			articliʒ́s	iʒ́s	VC∗
T1042	*happen*	vps	hapins	s	C
T764			happynīs	īs	C∗
T354			happynʒ́s	ʒ́s	C∗

(B) After

Text	Lexel	Grammel	Form	Inflection	String	IFS
T752	*article*	npl	articlis	is	VC	none
T496			articliīs	is	VC	horizontal stroke
T1316			articliʒ́s	is	VC	upward stroke
T1042	*happen*	vps	hapins	s	C	none
T764			happynīs	s	C	horizontal stroke
T354			happynʒ́s	s	C	upward stroke

other than <(V)C>. As with {S}, the low-frequency <VV> and <CC> tokens are discarded. There are two vpp tokens with initial <h> which are also discarded. The siglum-embellished tokens are merged with their no-siglum counterparts, and the sigla recorded as a separate variable for the token. Final <e>, which occurs in a total of 107 tokens, is also retained as a separate variable and the base form without the <e> merged with the corresponding form. Some examples of the recategorisation of {S} and {D} tokens with sigla or final <e> are shown in table 4.8.

After these omissions and mergers have been actioned, the potential orthographic realisations of OSc {S} and {D} can be represented in reduced form as:

{S} → ſ / **zero** / **(V)C**

{D} → **zero** / **(V)C**

It is then possible to categorise each token according to the orthographic form of its inflection using a manageable number of categories. <ſ> and zero are classified as distinct categories, and the remaining tokens are categorised according to their vowel and consonant components. Table 4.6 shows the potential values of the V and C characters and their attested combinations for each of the grammels npl, vps, vpp and vpt.

This categorisation structure means that the inflection of each token in the dataset is assigned a value for each of the following DVs:

(a) Zero

A binary variable with values 1 (inflectional form IS zero) and 0 (inflectional form IS NOT zero)

(b) Abbreviation

A binary variable with values 1 (abbreviation) and 0 (not abbreviation).

(c) Vowel

A categorical variable with values 'e', 'i', 'y' and 0 (no vowel).

(d) Consonant

A categorical variable with values 's', 'ʒ' and 'ß' for {S}; and '^t', 't' and 'd' for {D}.

TABLE 4.6: Potential values of V and C characters in {S} and {D} and their attested combinations. NA = not attested.

(A) npl

	ß	**s**	**z**	**Total**
e	eß (15)	es (88)	ez (431)	534
i	iß (37)	is (2,482)	iz (3)	2,522
y	yß (33)	ys (528)	NA	561
0	ß (9)	s (333)	z (78)	420
Total	**94**	**3,431**	**512**	**4,037**

(B) vps

	ß	**s**	**z**	**Total**
e	eß (2)	es (21)	ez (10)	33
i	iß (9)	is (637)	NA	646
y	yß (13)	ys (228)	NA	241
0	ß (8)	s (16)	z (1)	25
Total	**32**	**902**	**11**	**1,961**

(C) vpp

	^t	**d**	**t**	**Total**
e	NA	ed (16)	et (24)	40
i	i^t (1)	id (24)	it (2,812)	2837
y	y^t (10)	yd (20)	yt (777)	807
0	^t (171)	d (212)	t (365)	748
Total	**182**	**272**	**3,978**	**4,432**

(D) vpt

	^t	**d**	**t**	**Total**
e	e^t (1)	ed (1)	et (5)	7
i	i^t (4)	id (2)	it (999)	1005
y	y^t (6)	yd (1)	yt (223)	230
0	^t (124)	d (15)	t (109)	248
Total	**135**	**19**	**1,336**	**1,490**

TABLE 4.7: A sample of the dataset tabulated according to the DVs ABBREVIATION (A), VOWEL (V) and CONSONANT (C).

Lexel	Grammel	Form	A	V	C
good	npl	gud#(is)	1	-	-
bind	vps	bynd#(is)	1	-	-
claim	npl	clam#eȝ	0	e	ȝ
oblige	vps	oblig#es	0	e	s
scathe	vpp	scath#et	0	e	t
burn	vpt	bre(n)n#ede	0	e	d
land	npl	land#iß	0	i	ß
give	vps	gev#is	0	i	s
poind	vpp	po(n)d#id	0	i	d
award	vpt	award#it	0	i	t
arm	npl	arm#yxs	0	y	s
bear	vps	ber#ys	0	y	s
warn	vpp	warn#yd"	0	y	d
ward	vpt	wardd#y^t	0	y	^t

TABLE 4.8: Some examples of the recategorisation of tokens with sigla (S) or final <e> (FE).

Token with LAOS sigla representation	String	V	C	S	FE
$time/npl_tym#e^s	V∗C	e	s	h.mark	0
$mail/npl_mal#eȝ	VC∗	e	ȝ	h.mark	0
$allegation/npl_allegacioun#s	C∗	0	s	h.mark	0
$sheriffdom/npl_s(er)adom#e"s	V∗C	e	s	up.mark	0
$term/npl_t(er)m#is"	VC∗	i	s	up.mark	0
$annexation/npl_a(n)nexacion#s"	C∗	0	s	up.mark	0
$name/npl_nam#yxs	V∗C	y	s	v.mark	0
$happen/vps_happyn#s	C∗	0	s	h.mark	0
$happen/vps_happyn#s"	C∗	0	s	up.mark	0
$fasten/vps_festn#yxs	V∗C	y	s	v.mark	0
$ken/vpp_ke(n)n#it	VC∗	i	t	h.mark	0
$affix/vpp_affix#t"	C∗	0	t	up.mark	0
$touch/vpp_touch#it"	VC∗	i	t	up.mark	0
$ming/vpp_me(n)g#yxd	V∗C	y	d	v.mark	0
$ken/vpp_ken#de	Ce	0	d	0	1
$call/vpp_call#ite	VCe	i	t	0	1
$prove/vpt_pruf#it"	VC∗	i	t	up.mark	0
$gar/vpt_ger#t"	C∗	0	t	up.mark	0
$till/vpt_tel#de	Ce	0	d	0	1
$burn/vpt_bre(n)n#ede	VCe	e	d	0	1

4.1.5 Predictor variables

Having discussed and refined the DVs describing the realisation of the OSc {S} and {D} morphemes in section 4.1.4, in this section I introduce the predictor variables (PVs). That is, the factors which are being investigated as having a potential effect on the DV. In section 4.1.5.1, I describe the variables which pertain to lexical aspects (henceforth *lexical* PVs). I show their overall distribution in the corpus by quantifying each level and showing how they relate to one another. This is an important step in understanding potential

collinearities in the data. That is, the potential for two predictors to be correlated with one another, making it appear that both are correlated with the DV when in fact only one is actually correlated. I then describe the PVs which encode information relating to the source texts. I give an overview of the period and geographic area covered by the data, as well as how many data there are representing different times and spaces. I show the various text types represented in the corpus; how many texts and tokens represent each type, and in which time periods and geographical areas they occur. Textual data are extracted from the LAOS IoS (Williamson 2008), and will henceforth be referred to as *contextual* PVs.

4.1.5.1 Lexical PVs

In most cases, the lexels assigned in LAOS are directly representative of the word they represent in its uninflected form. For example, $land indicates the singular noun *land* when tagged with the grammel n, and the plural *lands* when tagged with npl. Both of these are completely predictable from the combination of the lexel and grammel which comprises the token's tag. For the majority of tokens, then, the terms *lexel* and *lexeme* are equivalent. However, for some derived lexemes, the lexel indicates the source of the derivation, rather than the derived form. An example of this is $break/npl, which indicates the plural noun *breakers* rather than the plural noun *breaks* as might be expected based on the pattern of *land* and *lands*. This example of *breakers* demonstrates a tendency in LAOS to tag deverbal nouns with a lexel indicating the noun's verbal root instead of its nominal derivation. LAOS is generally consistent in this practice, though there are a small number of exceptions: the lexeme $breaker also exists, as do other {- ER} derived nominal lexels, such as $absenter and $arbitrator. In the interest of consistency in the lexel variable (such as different tokens of *breakers* being tagged as $break/npl and as $breaker/npl); and because this study requires the comparison of verbal and nominal forms, I manually checked the data for deverbal nouns and assigned an appropriate noun lexel to those which are tagged with a verb lexel in LAOS.

In addition, some homonymous lexels have a semantic specifier placed within braces, such as $lie{f} (denoting *lie* in the sense of '[to tell a] **f**alsehood') and $lie{p} (denoting a **p**ositional sense of *lie*). These semantic specifiers are not numerous, nor are they judged to be of importance to this investigation. They are therefore omitted from INFLAOS and the lexels recorded with only their base form.

For each lexel in LAOS, I assigned a value based on the number of syllables I judged the uninflected word form to typically contain in ModSc. As noted in the previous section, the lexel used to tag a specific token is not in all cases representative of the manuscript word form. The revised lexels were therefore used as the basis for assigning syllable counts. In addition to these lexels, there are others used to tag deverbal past participle verb forms prefixed with {UN-}, as in examples (15a) and (15b). In these cases, as there is no word-class change caused by the prefixation, and the immediate environment of the crucial inflectional suffix is not affected, I have retained the original LAOS lexel but altered the syllable count. Example (15a)

therefore has a syllable count value of 1, and example (15a) a value of 2.

(15) a. <harmit> vpp *harmed*

b. <vnharmit> vpp *harmed*

A potentially confounding factor for the categorisation of tokens by stem syllable count is the potential for the addition of an epenthetic vowel in certain phonetic contexts. This is incorporated into the analysis on a qualitative basis in section 10.3.1.

Each lexel is also assigned the value 'Germanic' or 'Non-Germanic'. These categorisations are drawn from the *Oxford English Dictionary* (*OED*) and the *Dictionary of the Scots Language* (*DSL*).

Each token is assigned a value according to the final character of the transcribed form of its stem, with the following exceptions:

(a) stems ending in a mark of suspension are assigned a value corresponding to the entire expanded form. For example, p(ert)i(nent)#(is) *pertinents* (npl) is assigned an stem-final *littera* (SFL) value '(nent)'

(b) stems ending in sigla are assigned a value combining the final alphabetic character and the siglum. For example, <som-#ys> *sums* (npl) is assigned an SFL value 'm-'.

(c) stems ending in an indecipherable character are omitted from the dataset (see Appendix A). For example, <hap[]#is> *happens* (vps).

(d) stems ending in a superscript letter are assigned a value indicating this. For example, <ry^t#(is)> is assigned an SFL value '^t'.

TABLE 4.9: The original categories of the SFL variable, and the frequency with which they occur in the dataset.

SFL	Tokens	SFL	Tokens	SFL	Tokens
d	4660	**f**	481	**i**	19
r	4579	**h**	409	**b**	15
t	3542	**(n)**	336	**a**	13
n	2921	**u**	251	**(is)**	12
s	1361	**p**	215	**o**	9
l	1337	**c**	189	**(ro)**	5
m	1215	**v**	174	**(ar)**	2
e	1129	**(ur)**	148	**(nent)**	1
(er)	888	**(e)**	124	**^d**	1
k	785	**x**	119	**^e**	1
g	779	**y**	68	**^h**	1
ß	590	**(m)**	32	**^n**	1
w	531	**z**	27		

Table 4.9 shows the resultant categories of the SFL variable, and the frequency with which they occur in the dataset. In total, there are 46 unique SFL attested in the dataset, 15 of which are represented by

five or fewer tokens. Particularly low-frequency categories are superscript *litterae*, suspension marks and vowel *litterae*. It is possible that these three types might form internally consistent groups, particularly the suspension marks, some of which are likely to be similar to one another regardless of the letter they represent (Simpson 1973). However, the low frequency of these tokens precludes any investigation to ascertain whether or not this is the case. Tokens with stem-final suspension marks are accordingly categorised the same way as tokens with stem-final *sigla* represented in LAOS as " and ~. Similarly, the low frequency of superscript SFL means that it is not possible to investigate whether these categories form a coherent group or whether there are grounds to group each superscript *littera* with its corresponding full *littera* (for example, final <d> with final <^d>). These tokens are accordingly also omitted from the dataset (see Appendix A). The final low frequency type, vowel SFL, are understandably rare in this dataset due to the omission of phonetically vowel-final lexemes discussed in section 4.1.2 and the reassignment of morpheme boundaries in tokens with stem-final <eC> discussed in section 4.1.3. Nonetheless, there are 141 tokens with a vowel as the SFL remaining in the dataset at this stage, the majority (92) of which have stem-final <e> followed by covered inflectional <i>. For the same reasons as I detailed in section 4.1.2, tokens which are orthographically vowel-final are judged to introduce ambiguity into the process of uniquely identifying inflection forms as distinct from their stems. They are therefore omitted from the dataset (see Appendix A). Table 4.10 shows the categories of the SFL variable once the superscript-, suspension- and vowel-final tokens have been omitted (see Appendix A).

TABLE 4.10: The revised categories of the SFL variable, and the frequency with which they occur in the dataset.

SFL	Tokens	SFL	Tokens
d	4,728	**u**	538
r	4,694	**w**	530
t	3,774	**f**	505
n	2,936	**h**	459
s	1,368	**y**	321
e	1,358	**p**	247
l	1,339	**o**	205
m	1,219	**c**	193
g	790	**v**	173
k	784	**x**	119
ß	590		

4.1.5.2 Contextual PVs

Each text has an index number which uniquely identifies it in LAOS. I include this unique identifier of individual texts as a PV to account for the variation in the data caused by individual text idiosyncrasies. I discuss this further in section 5.2.5.

The IoS lists the location of origin of each manuscript where it is stated within the text itself. As

the documents are legal records, the place of issue or signature is often recorded within the text itself, so the majority of the texts in the corpus have a location specified. These locations have been converted into British National Grid coordinates in the IoS, and into geographic coordinates (latitude and longitude) in INFLAOS using an online public-access co-ordinate conversion tool (Ordnance Survey Ltd. 2016). Some texts are not assigned coordinate values in the IoS. Where this is the case, the text and all tokens associated with it are retained in the dataset but given null values for the variables latitude and longitude.

As with text location, the legal status of the majority of the LAOS manuscripts meant that they were marked with the date of composition or issue within the text. Where it exists, this information is included in the IoS. I have included in my analysis the year of composition of each text, disregarding the day and month if mentioned. Where a period of years is specified, I take the later year as representative of the text (for example, a text with an entry in LAOS specifying a date range of 1477 to 1478 is assigned the date 1478).

The majority of texts in the IoS are assigned a type. Some have a single-category type (examples 16a and 16b). Most have at least one subtype (example 16c), with some texts having as many as four subtypes (example 16d).

(16) Examples of LAOS text type categorisations:

a. Text 1415: **cartulary**

b. Text 93: **charter**

c. Text 238: **cartulary**/ copy/ lease

d. Text 960: **charter**/ transumpt/ signet letter/ exchange/ excambion

Taking into account all the permutations of different text categories specified in the IoS, there are a total of 188 unique classifications. Clearly, it is not practical or, for that matter, necessary to retain all details of the original text type categorisation for the purpose of statistical anlysis. The five most common primary categories (those listed in first position in the IoS) in my dataset are listed in table 4.11a. These five categories together account for 1,097 texts of a total 1,199 in my dataset (91%). 12 texts (1%) are split between the primary categories listed in 4.11b.

84 texts (7%) are not assigned a type in the IoS. However, 62 of these are noted as being transcribed from a manuscript which is clearly a burgh record book: 54 texts numbered 1801 to 1861, as well as text 3007, are all transcribed from the *Burgh Court Book of Dunfermline*; five texts numbered 3001 to 3006 are transcribed from the *Aberdeen Council Register*; and text 3008 is transcribed from the *Ayr Burgh Court Book*. Based on this information, I assigned the type 'Burgh Record' to these texts. The remaining 12 texts and their tokens are retained in the dataset but assigned a null value for the variable of text type.

(A) Most common text type primary categories.

Category	Texts
Book	482
Charter	343
Cartulary	110
Notarial Protocol Book	108
Burgh Record	54
Total	**1,043**

(B) Least common text type primary categories.

Category	Texts
Summons	5
Letter	2
Record	2
Assize	1
Bond	1
Court Book	1
Credence	1
Deed Poll	1
Inscription	1
Record Book	1
Royal Letter	1
Signet Letter	1
Total	**12**

Reducing the list of text types to the primary categories named in the IoS yields the 17 categories shown in tables 4.11a and 4.11b. However, some category names suggest a high potential for textual similarity. For example, the separate categories 'Burgh Record' and 'Record Book' seem, *prima facie*, likely to have a great deal in common. This is not to say that a differentiation as it is presented in the IoS is unwarranted, but rather that for the purposes of my analysis, there is no reason to suppose that such fine-tuned categorisation is necessary. In the case of 'Burgh Record' (54 texts) and 'Record Book' (one text), consultation of the IoS shows that all four texts are extracts from the *Peebles Burgh Council Record Book*. Though it is entirely possible that a thorough reading of and research around these texts such as that undertaken by Williamson (2008) in compiling LAOS would uncover a nuanced difference behind these slightly different types, the fact that the texts in question come from the same council record book suggests that the categories of 'Burgh Record' and 'Record Book' can be combined, certainly for the purposes of my analysis, into one category, taking the title of the largest category, 'Burgh Record'.

The previous example is clear-cut, however, it is not always possible to make concatenation decisions based on primary categories. The type 'Record' (two texts) sounds like a potential candidate for inclusion under 'Burgh Record'. However, the subcategories of these two texts are 'assize' and 'ecclesiastical'. Further examination of the former shows that it is more appropriately categorised alongside the text which has the primary category 'Assize'. Furthermore, there are five other texts which have the subcategory 'Assize' and which all have the primary category 'Charter', suggesting that it may be possible to merge the types 'assize', 'record/assize' and 'charter/assize'.

Having examined the text typology in the IoS together with the text content, I ultimately reduce the text type categorisation in my dataset to the groupings shown in table 4.12, which specifies the overarching type I use in this study as well as the LAOS primary categories which have been merged to create each category. The table also specifies a short code for each text type for ease of reference and presentation.

Two text types warrant further explanation. Firstly, the text type 'inscription' (listed in table 4.11b) is represented by only one text (text 97, inscription on a memorial stone, ca. 1380) which contains only three INFLAOS tokens. This text (and consequently, type) has therefore been omitted from the dataset (see Appendix A). Secondly, the text type 'State Documents' (STA) in table 4.12 is the only type category label which does not take its name from a text type category used in LAOS. Rather, this text type grouping consists of state documents which have a code beginning with 'X' in place of a county label: 'XAP' indicates an Act of Parliament; 'XDI' indicates a diplomatic treaty; and 'XST' indicates an unspecified type of state document. Texts with the county code 'XLC' are not state documents, but rather unlocalised texts.

TABLE 4.12: Text type categorisations used in this study and the LAOS primary categories they represent.

Label	Contents	LAOS Types	Texts
BUR	Burgh or council records.	Book	482
		Burgh Record	116
		Record Book	1
		Court Book	1
		Total	**600**
CHA	Charters - this is a somewhat catch-all type, essentially covering any legal documents which do not belong to the other categories.	Charter	343
		Letter	2
		Record*	2
		Assize	1
		Bond	1
		Credence	1
		Deed Poll	1
		Royal Letter	1
		Signet Letter	1
		Total	**353**
CAR	Cartularies - this category can be generally assumed to contain monastery copies of charters.	Cartulary	110
		Total	**110**
NOT	Notarial protocol books. According to the *Dictionary of the Scots Language* (DSL), "a book or register kept by a notary, and containing records of transactions or other legal proceedings attested by him in his official capacity." (*Prot(h)ogoll, -col(l, n.* 2004)	Notarial Protocol Book	108

TABLE 4.12: Text type categorisations used in this study and the LAOS primary categories they represent.

Label	Contents	LAOS Types	Texts
		Total	**108**
STA	State documents - these are texts which have a county code which, rather than specifying a county, identifies the text as a state document: an act of parliament (XAP); a diplomatic treaty (XDI) or an unspecified type of state document (XST).	Charter	34
		Summons	5
		Total	**39**

Figure 4.5 shows the distribution of INFLAOS texts across the time period covered by LAOS, 1380 to 1500. The graph shows that the number of texts representing each decade is comparatively low in the first 40 years of the period, but rises sharply between 1430 and 1460, the period to which 478 of the 1,217 texts in INFLAOS (approximately 40%) are dated. By contrast, only 87 texts (7%) are dated earlier than 1430.

Figure 4.6 shows the distribution of INFLAOS tokens across the same period. The line of this graph follows a smoother upward curve than that showing the distribution of texts. Whilst it shows an increase in tokens after 1430, there is not the dramatic spike shown by the increase in the number of texts, nor is there the subsequent drop in numbers after 1460 which the graph of text counts shows. This indicates that, whilst there is a sharp rise in the number of texts after 1430, this rise may be characterised by an increase in shorter texts, explaining why the token count of INFLAOS does not spike concomitantly with the text count.

Figure 4.7 supports this conclusion. It shows the average number of tokens per text for each decade. As suggested by figures 4.5 and 4.6, the beginning of the period is characterised by a high average number of tokens per text, whereas the middle of the period shows this average dropping substantially.

Figure 4.8 suggests a reason for this. It shows the number of texts of each type per decade in INFLAOS. The overall shape of the bars follows a similar pattern to the shape of the line in figure 4.5, and it is clear

that this shape is defined by one particular text type - Burgh Record. It also shows that the second increase in volume of texts shown by figure 4.5 is largely due to the appearance of texts from Notarial Protocol Books. There are three notaries public named in LAOS whose records of legal transactions are included in the corpus. All three of these books are dated after 1480.

Table 4.13 lists the total text and token counts for each text type, as well as the average number of tokens per text of each type. Burgh records have the lowest average number of tokens per text and the highest overall total of texts. The table also reveals that INFLAOS contains the same number of cartularies and notarial protocol books. These two text types contribute a similar number of overall words to the corpus and have a similar average number of words per text. However, as shown by figure 4.5, cartularies appear throughout the period, whereas notarial protocol books appear only at the very end.

TABLE 4.13: The distribution of INFLAOS texts and tokens by text type.

Type	**Total texts**	**Total tokens**	**Avg. tokens/ text**
BUR: Burgh Record	620 (51%)	6,452 (25%)	10
CHA: Charter	301 (25%)	10,412 (40%)	35
CAR: Cartulary	110 (9%)	2,663 (10%)	24
NOT: Notarial Protocol Book	110 (9%)	2,376 (9%)	22
STA: State Document	56 (5%)	3,048 (12%)	54
NA: no type specified	20 (2%)	763 (3%)	38
Total	**1,217**	**25,714**	

The distribution of texts and tokens is inconsistent over time, as is to be expected with opportunistic corpus data. It is also inconsistent with regard to text types and lengths, with 51% of texts being identified as burgh records, but 40% of tokens coming from texts identified more generally as charters.

The temporal distribution of the different text types identified in LAOS is shown together with their geographical distribution in figure 4.9. Each map contains circular markers representing the texts in INFLAOS within a 20-year period. The colour of the marker denotes the type of text represented, and the size of the marker denotes the number of texts represented. The first two maps show the comparatively small number of texts attested in the first 40 years of the corpus, and the fact that these texts are mainly charters (CHA) and state documents (STA). These two observations have already been noted in the previous discussion, but the visualisation of these data in their geographical context shows that in the earliest years of the period, texts were centred mainly around Edinburgh and the Lothians, with only a few texts located in peripheral areas. Throughout the period, texts continue to be clustered most densely around Edinburgh and the surrounding areas though, after 1440, this area of density extends further up the east coast towards Aberdeen. In the period 1460-1500, texts also begin to be attested in greater number towards the west coast. In terms of text type, I have already noted the difference in temporal distribution between charters and burgh records, but the maps in figure 4.9 illustrate most clearly how the burgh records in the

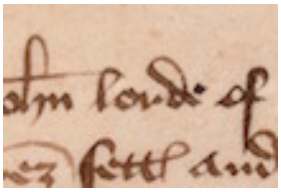
corpus occur in sudden, large pockets throughout the period. For example, the large, red marker located at Aberdeen on the 1440-1460 map represents the texts transcribed from the *Aberdeen Council Register*, and the larger of the two red markers on the 1460-1480 map represents texts transcribed from the *Burgh Court Book of Dunfermline*. Other observations which can be drawn from these maps are:

(a) Notarial protocol books (NOT) are, like burgh records, attested in bursts at a particular time and location. Unlike burgh records, however, notarial protocol books are attested in only two locations extremely close to one another (Edinburgh and Peebles), and all within the same 20-year period, 1480-1500.

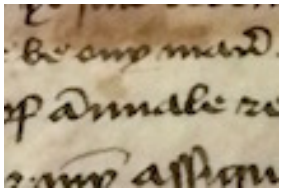
(b) Cartularies (CAR) differ from burgh records and notarial protocol books in that they are fairly evenly distributed over time. However, figure 4.9 shows that cartularies are most often attested within a particular area, specifically the counties of Fife and Angus.

(c) The level of detail in the coordinates given for particular texts, whilst useful in many ways, can sometimes obscure broader trends in the data. For example, the 1480-1500 map shows two effectively superimposed markers for state documents localised to Edinburgh. The text coordinates which are at the centre of each marker are different, but similar enough that the markers completely overlap. This is also often the case with markers representing charters - there are many charters which are localised to points very close to one another, but not exactly the same. The consequent visual effect is to diminish the number of charters represented by smaller, overlapping points, in comparison to the groups of texts which are necessarily assigned identical coordinate values, such as the records of a single burgh court.

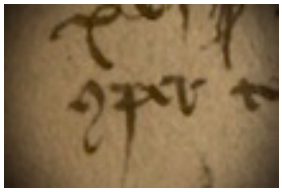
FIGURE 4.2: Examples illustrating LAOS transcription conventions from text 36 [1447, CHA, ELO]; text 40 [1426, CHA, AYR]; text 1608 [1436, BUR, ABD]; text 1621 [1443, BUR, ABD]; and text 1644 [1445, BUR, ABD]. (D) and (F) are reproduced from Johnson and Jenkinson (1915: 56) due to the lack of available manuscript images including them.



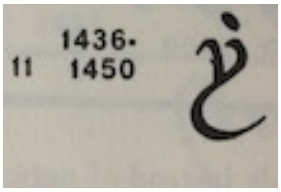
(A) LORDE (T36)
lord



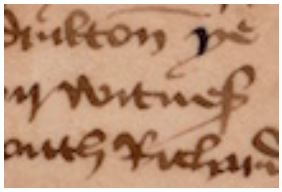
(B) AnNUALE (T40)
annual



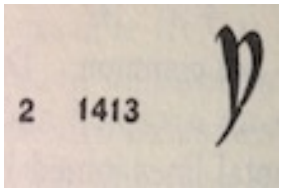
(C) comPER (T1644)
compear



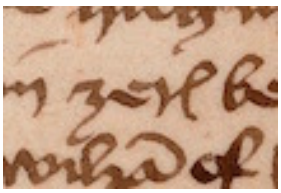
(D) Yx
(<y> with diacritic)



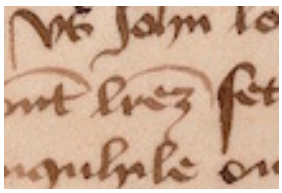
(E) WITNESs (T36)
witness



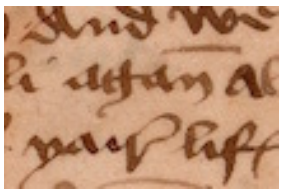
(F) y
(y/þ character)



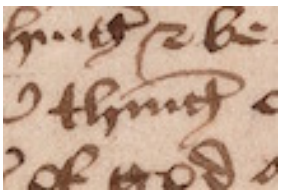
(G) zER+is (T36)
years



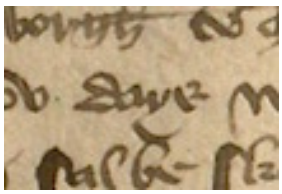
(H) LettREz (T36)
letters



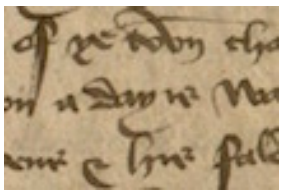
(I) agan- (T36)
against



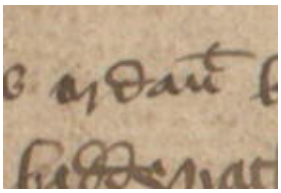
(J) THING" (T36)
thing



(K) DAY+S (T1621)
days

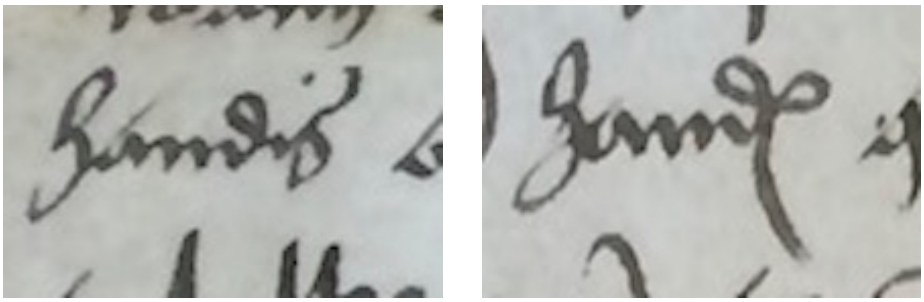


(L) DAY-IS (T1621)
days



(M) ORDAN+^T (T1608)
ordained

FIGURE 4.3: Two instances of $hand/npl, both from text 7 [1493, CHA, AYR].



(A) HAND+IS
hands

(B) HAND+is
hands

FIGURE 4.4: The manuscript abbreviation for <(us)>, reproduced from Cappelli (1982: 13).

FIGURE 4.5: The number of texts in INFLAOS by decade.

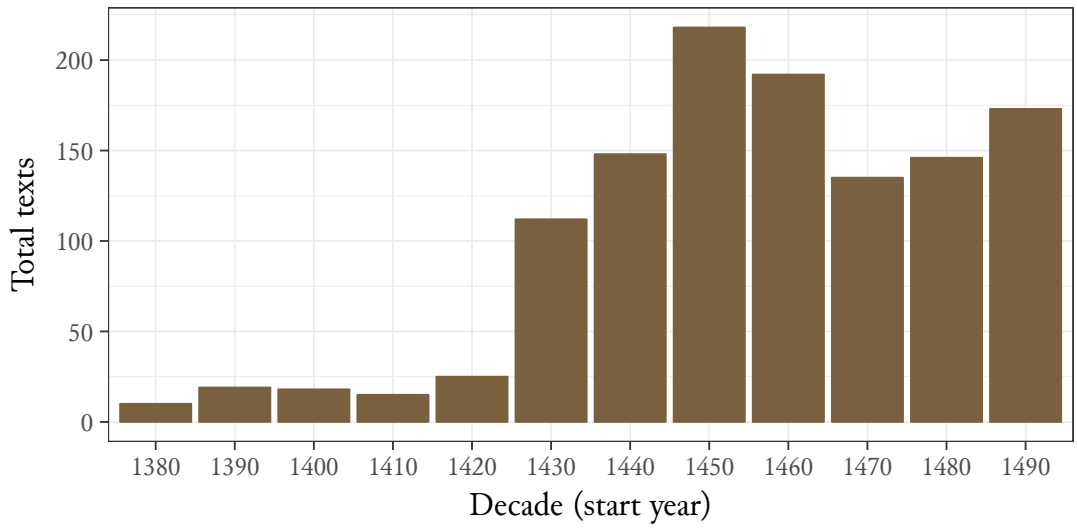


FIGURE 4.6: The number of tokens in INFLAOS by decade.

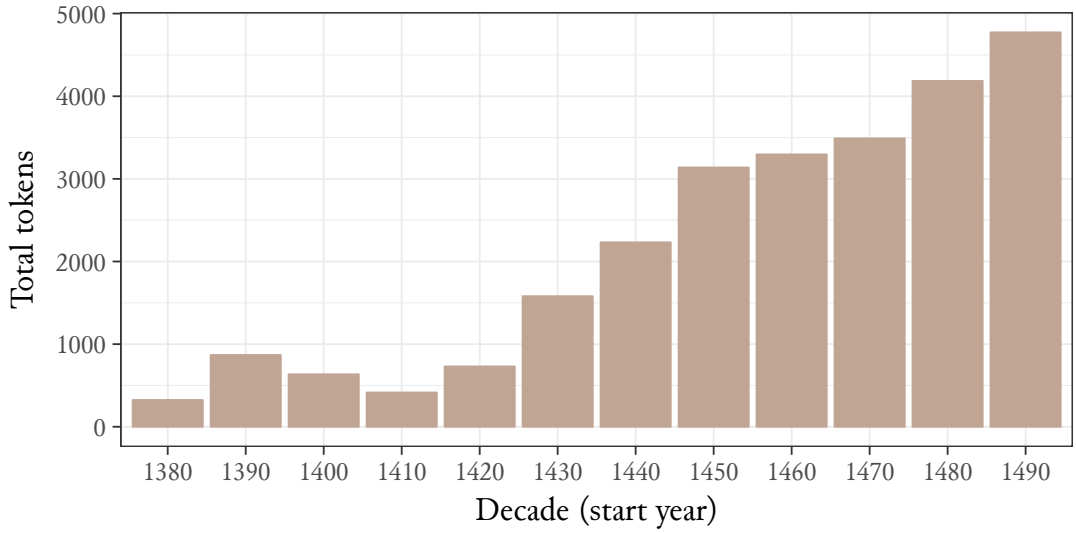


FIGURE 4.7: The average number of tokens per INFLAOS text by decade.

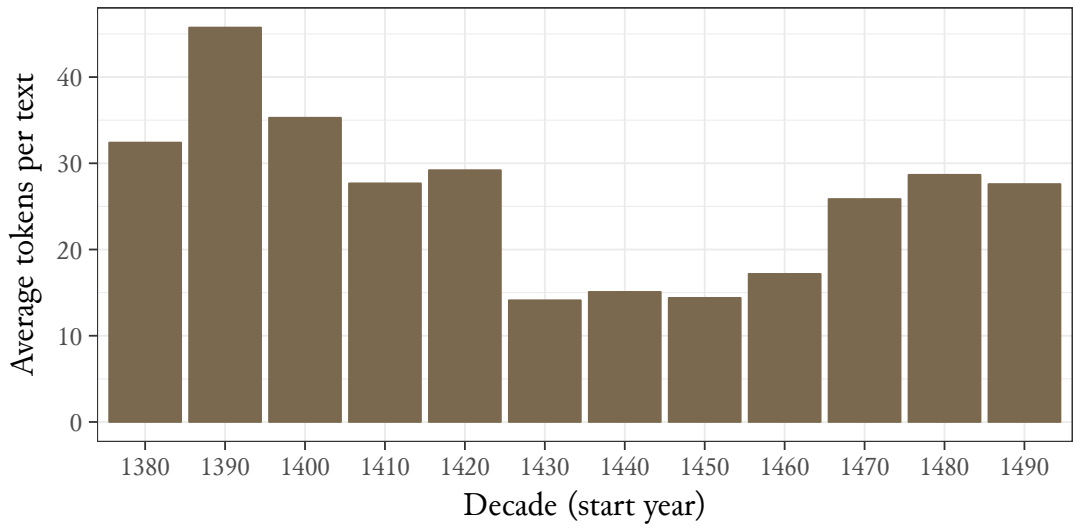
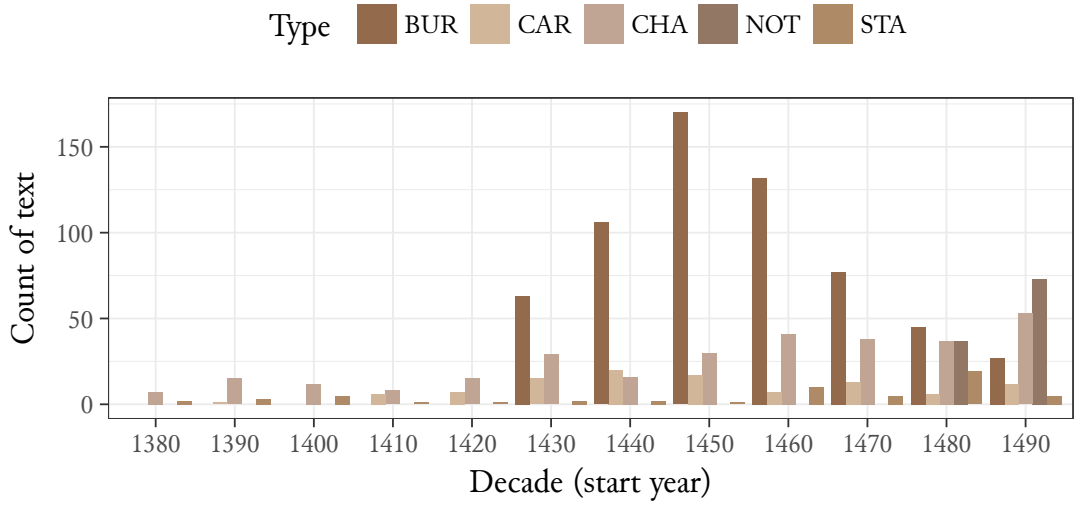


FIGURE 4.8: The number of texts of each type in INFLAOS by decade. Types are BUR (burgh record); CAR (cartulary); CHA (charter); NOT (notarial protocol book); and STA (state document). For an explanation of each text type, see section 4.1.5.2.



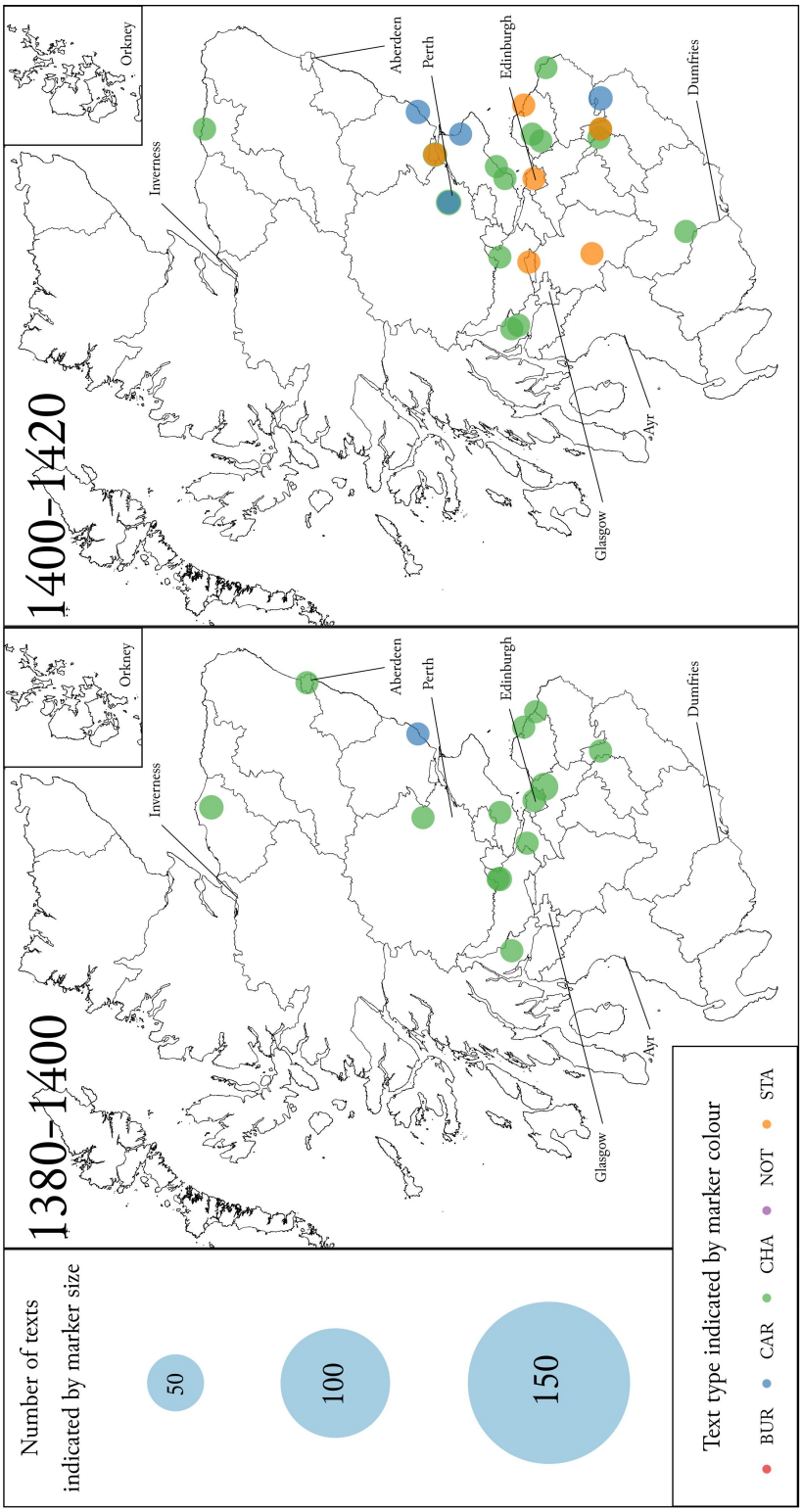


FIGURE 4.9: Distribution of INFLAOS texts of different types in the periods 1380–1400 and 1400–1420

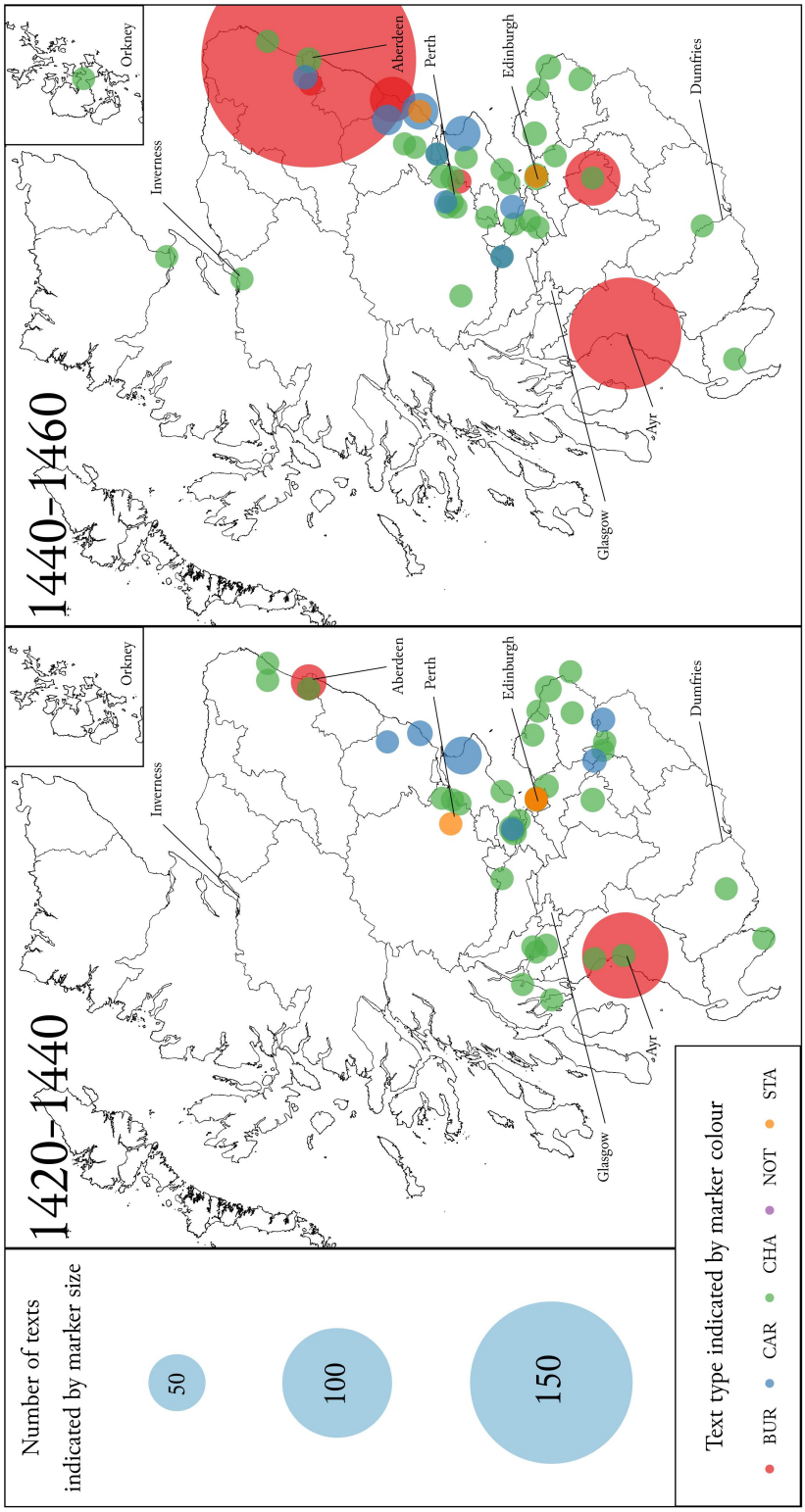


FIGURE 4.9: Distribution of INFLAOS texts of different types in the periods 1420–1440 and 1440–1460

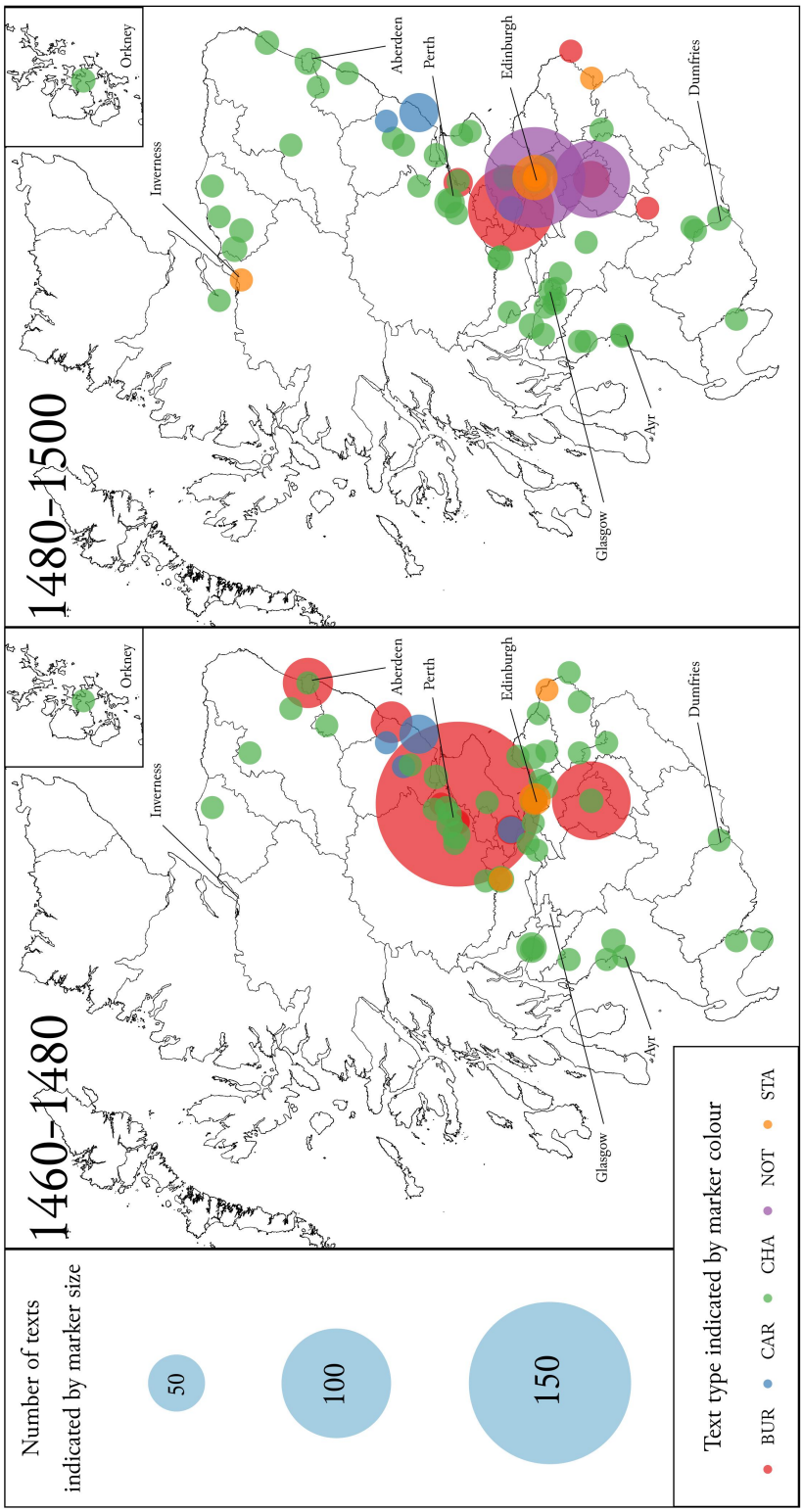


FIGURE 4.9: Distribution of INFLAOS texts of different types in the periods 1460–1480 and 1480–1500

Chapter 5

Statistical Methods

I aim in this section to provide an overview of the statistical methodology I use to investigate the realisation of the Older Scots (OSc) {S} and {D}, specifically, generalised additive model (GAM). This overview is aimed primarily at historical linguists, not at statisticians, and consequently I attempt to simplify my explanation. The mathematical theory underlying this methodology is extremely complex and, for in-depth explanation, I refer the reader to texts which have been instructive during my study of these techniques: Wood (2006); Baayen (2008); Zuur (2009); Wieling et al. (2014); and Wieling and Nerbonne (2015). In particular, Wood (2006) is the primary source for the mathematical theory underlying the concept of GAM. Wieling and Nerbonne (2015) demonstrate the advantages of GAM analysis from a (socio-)linguistic perspective.

In section 5.2, I introduce the concept of linear modelling by presenting a simple example. The example is trivial in itself, but exemplifies the concept of linear regression using the *Inflections in A Linguistic Atlas of Older Scots* (INFLAOS) dataset. In section 5.2.1, I present another example of an linear model (LM), this time a generalised linear model (GLM), which allows for the modelling of binomially distributed data (where the dependent variable (DV) takes one of two possible values) as opposed to continuous data. The plots of the LM and the GLM constructed in these first two sections are compared, showing that in the simple LM, the values of the continuous DV are modelled directly according to the relationship between the DV and the predictor variable (PV), whereas in the GLM, it is the *likelihood* of one level of the binary DV as opposed to the other which is modelled according to this relationship.

In section 5.2.2, I point out the limitations of the (G)LM models introduced in sections 5.2 and 5.2.1. Specifically, the fact that GLMs assume a linear relationship between the PV and DV. This is not suitable for modelling data which varies unevenly over time and space (i.e. a correlation which cannot be graphed as a straight line) such as INFLAOS. In the previous sections, I do not address the spatio-temporal PVs in INFLAOS, however in this section I present a comparison of the same subset of INFLAOS modelled using

(a) a GLM and (b) a generalised additive model (GAM). Using plots of the model results, I demonstrate that the ability of the GAM to account for non-linear relationships between the DV and PV using smoothing techniques enables far superior insight into the patterns shown by the INFLAOS data.

Section 5.2.4 expands on the single-PV models of section 5.2.2, demonstrating the modelling of the effect of interactions between PVs as well as their individual effects. I demonstrate that whilst the GLM reveals a difference between the relationship of Germanic and Non-Germanic lexis with the DV over time, it is only able to estimate a linear regression line for each level. In other words, only the overall correlation between the DV and the PV of time can be shown, not how the magnitude or direction of this correlation varies over time. The GAM, on the other hand, reveals a far more complex picture of the relationships between these variables.

Section 5.2.5 focusses only on GAM, examining in more detail the results of the GAM in section 5.2.4. Specifically, this section introduces the concept of *random effect* PVs, showing how individual variation at the level of lexel groups can be misidentified as correlation with a *fixed effect* PV if this variation is not incorporated into the model.

5.1 Some notes on terminology

5.1.1 Variants and variables

This thesis makes extensive use of statistical methods and terminology. It is therefore worthwhile to make some comments clarifying similar terms which are commonly used in both statistical and linguistic fields. The terms *variant* and *variable* are an example of this. In linguistics, it is common to refer to a *variant* - a realisation of a particular linguistic structure. For example, the {S} morpheme in OSc has several orthographic variants, such as <is>, <ys>, <ſ> etc. Similarly, in sociolinguistics, reference is often made to pronunciation variants, for example, the presence of absence of rhoticity in different English varieties.

The statistical concept of *variable* is closely related to that of *variant* when it comes to linguistic analysis. A *variable* is a linguistic structure which can have more than one *variant* or realisation. In some cases, it is necessary to analyse particular variables by categorising their variant forms in a way which is conducive to quantitative analysis. For example, it is not feasible to analyse the occurrence of all possible variants of OSc {S} as they are transcribed in *A Linguistic Atlas of Older Scots* (LAOS). Rather, it is necessary to group variant forms into manageable categories. There are many variant forms of {S} recorded in LAOS, but, as outlined in section 4.1.4, these variants have been categorised into binomial DVs in order to achieve a meaningful analysis of their distribution.

5.1.2 Effect sizes

Throughout this investigation, it will often be necessary to describe and to quantify the *effect* of a PV on a DV. It is important to clarify at the outset what is meant by this, as the implication of this term when used in reference to regression modelling is different to its common use.

When a regression model is fitted, what is being measured is the change in observed values of the DV at different levels of the PV. For example, a model of weight as a function of height measures the change in the value of weight at different values of height. This is referred to as the *effect* of height on weight, meaning the effect of a change in the PV height on the observed values of weight.

The characterisation of this relationship as an *effect* refers to the statistical model of the data - what is the *effect* on the value of weight if the value of height is increased from, say, five feet to five feet and three inches? This concept of *effect* is used to quantify the significance of a PV on a DV. In this case, does changing the observed value of height have a significant effect on the observed value of weight?

However, this does not imply a *causal* relationship between the two variables themselves. Rather, a regression model takes as its input the observed values of the DV and PVs. In regression modelling terms, this is referred to as the effect of the value of a PV on the value of a DV. A regression model estimates the overall effect that a change in PV value has on DV value and returns a significance estimate.

5.2 Linear modelling

GAMs are an extension of GLMs, which are in turn an extension of LMs. An LM is a representation of a numeric response variable modelled as a function of one or more PVs, plus an *error term* which specifies the variation in the observed data assumed to be due to random error. The more variation in the observed data that a model can account for with the PVs included in it, and consequently the less variation ascribed to random error (that is, variation in the data which is not captured by the PVs), the better the *fit* of the model. The better the fit of the model, the better it explains the observed data.

A *simple* LM is one which contains only one predictor variable. As an example, consider the variable syllable count in INFLAOS, and the below (deliberately self-evident) hypotheses.

> H_1: As syllable count increases, the number of orthographic characters used to represent a word also increases.
>
> H_0: There is no relationship between syllable count and the number of orthographic characters used to represent a word.

To test whether there is a relationship between syllable count and the number of orthographic characters used to represent a word, a simple LM can be constructed with the formula **number of characters ~ number of syllables**, where the tilde ~ between the DV on the left and the PV on the right means *predicted by*. The

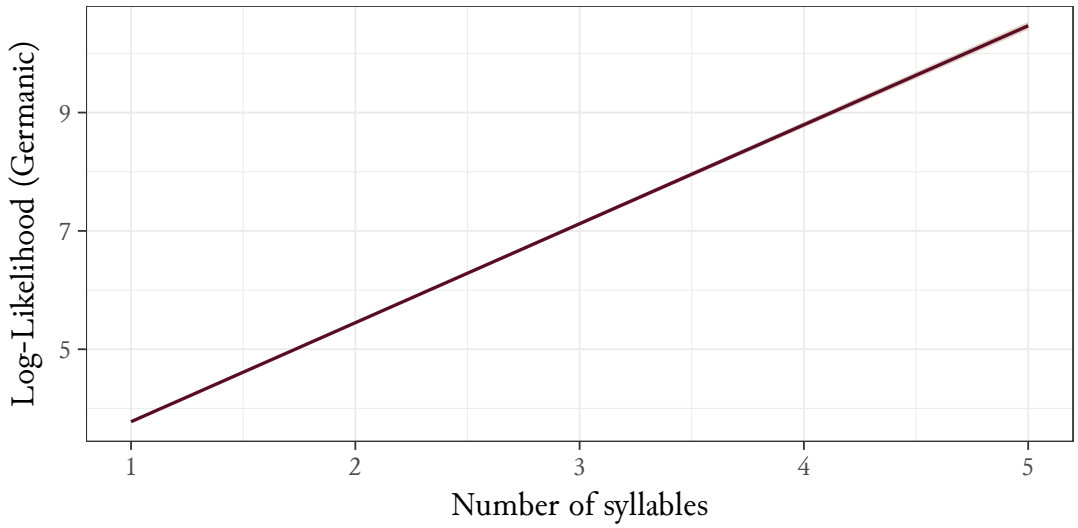
stated formula, then, can be read as "number of characters *predicted by* number of syllables". That is, the model attempts to use the PV (number of syllables) to predict the DV (number of characters). How well this can be accomplished gives us an indication of whether the two variables are significantly correlated with one another.

The simple LM calculates the relationship between the PV and the DV as follows:

(a) the observed data values are plotted as points with the DV on the y-axis and the PV on the x-axis as shown in figure 5.1.[1] Each point in the plot represents a data value, in this case, an INFLAOS token.

(b) A regression line is plotted through the points. In a simple LM with one PV, this is also known as a *line of best fit*. The position of the line is calculated to divide the data points into two equal groups above and below it. The gradient of the line indicates the direction and strength of the correlation between the DV and the PV. In figure 5.1, the line dividing the points into two groups has a fairly steep gradient and shows a positive correlation, meaning that the higher the value of the PV number of syllables, the higher the corresponding value of the DV number of characters.

FIGURE 5.1: The regression line produced by an LM which attempts to use the PV number of syllables to predict the DV number of characters



[1]Graphical representations of regression model results throughout this paper were created using R. Specifically, package ggplot2 (Wickham 2009) was used to visualise the output of package visreg (Breheny and Burchett 2016), and the resulting figures typeset using package tikzDevice (Sharpsteen and Bracken 2016).

5.2.1 Generalised linear modelling

A generalised linear model (GLM) is an LM which is extended to allow the DV to have a non-normal distribution. A normal distribution is characteristically represented by a bell-curve - most data points cluster towards the middle of the distribution, with steadily fewer data points symmetrically occupying the upper and lower extremes of the scale. A characteristic example is the heights of a population. Say the mean height of a population is five feet and six inches. Most people are clustered fairly close to the mean, with an approximately equal number of individuals being taller than average as those who are shorter than average. The vast majority of people are clustered around the average height, say between five and six feet tall. Then there is a minority of people who are much shorter or much taller than average. Following the logic of the normal distribution, the numbers of extremely short and extremely tall people will be approximately equal.

A non-normal distribution, then, is one which does not follow this pattern. There are many kinds of non-normal distribution, such as a skewed distribution, which is like a normal distribution but non-symmetrical, with more data points clustering either to the left or the right of the median value. A good example of this is global individual wealth distribution, where the vast majority of wealth is held by a small number of individuals.

In the following investigations, the most important distributions will be the binomial and logistic distributions. The binomial distribution describes a variable which can have one of two possible values. An example is the variable etymology in INFLAOS[2]. etymology is a binary variable with the values 'Germanic' and 'Non-Germanic'. Using the same continuous PV as in the LM example, number of syllables, a GLM can be specified with the formula **etymology ~ syllables**, meaning that the model attempts to predict the etymology of a token based on number of syllables. As with the LM, how well the model with these parameters fits the observed data indicates how correlated the DV and the PV are.

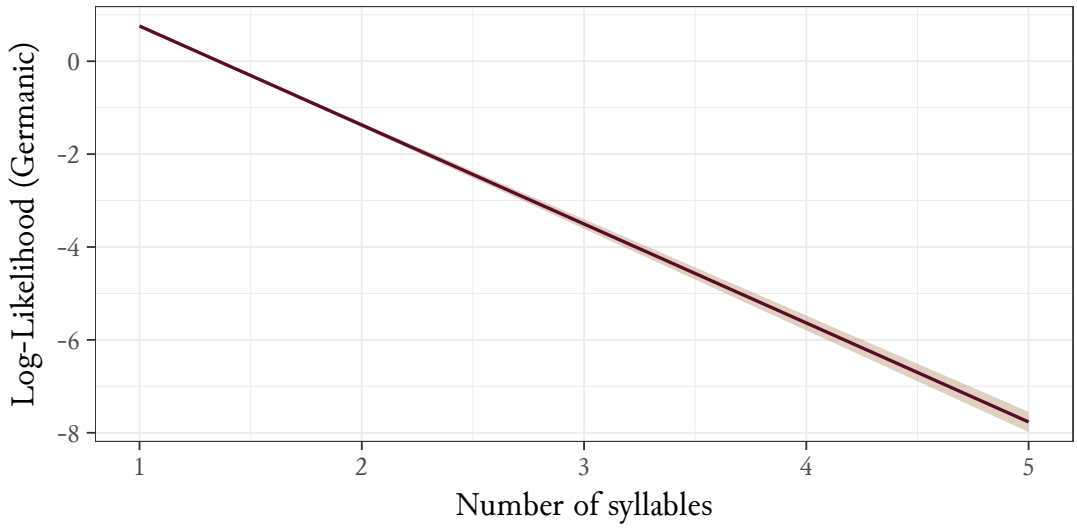
However, there is an added complication with the GLM formula compared to the LM. Namely, the LM could predict a numerical value of the continuous DV (orthographic characters) for a given number of syllables, but in this case, the DV has only two possible values. To understand the issue here, imagine trying to fit a regression line as described above for binomially distributed data points. The solution is to model the *log-likelihood* of a given level of the DV as predicted by the PV. The *likelihood* of a particular value of the DV can be conceptualised similarly to the *probability* of that value. We could pick a random token from INFLAOS and talk about the probability of it representing a Germanic lexel based on our knowledge of the internal composition of INFLAOS. The likelihood, on the other hand, is a value which is estimated by the GLM by observing all tokens in INFLAOS and their distributions, and deciding on a likelihood

[2]The general term for regression in which the DV follows a binary distribution is *logistic regression*. Logistic regression has been used extensively in sociolinguistics (Speelman 2014: 1), where it is often referred to as *Variable Rule Analysis* (VARBRUL). VARBRUL differs from GLM in that the effect of the PVs on the DV is estimated under the assumption that all PVs are independent. GLM on the other hand estimates the effect of each PV whilst taking into account the effects of other PVs in the model (Gries 2003: 155).

value which makes the observed data most probable.

The regression line plot for the GLM is shown in figure 5.2. Whereas the y-axis in figure 5.1 shows the values of the continuous dependent variable (number of orthographic characters), the y-axis in figure 5.2 shows the *log-likelihood* of a token having the *non-default* value of the DV. In this case, 'etymology = Germanic' is assumed to be the default value. The default value is essentially arbitrary, but where the values of a PV are discernibly positive and negative ('Germanic = yes' as opposed to 'Germanic = no'), so the regression line shown in figure 5.2 shows a positive correlation between syllable count and the likelihood that 'etymology = Non-Germanic'. That is, the model predicts that words with more syllables are more likely to be Non-Germanic than words with fewer syllables. This does not imply that there is a cause-and-effect relationship between these two variables, but rather that based on the observed data, there is a statistically significant reason to predict that, for example, a given 4-syllable word is likely to be non-Germanic.

FIGURE 5.2: The output of a GLM indicating a positive correlation between syllable count and the likelihood that 'etymology = Non-Germanic'.



5.2.2 Generalised additive modelling

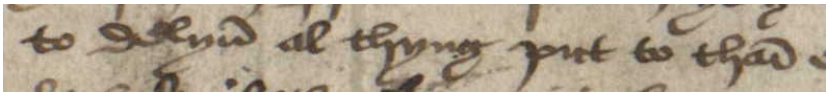
A crucial assumption made by the LM and GLM models in sections 5.2 and 5.2.1 is that the relationship between the response and PVs is linear (i.e. the correlation can be plotted as a straight line). If this is not the case (because the response variable is affected in different ways at different levels of the predictor), then linear models will not fit the data well. GAMs allow for the relationship between DVs and PVs to be non-linear. In other words, the relationship between the two variables can be plotted as a wiggly line rather than a straight one, thereby capturing a non-linear relationship between them more accurately. This

approach is referred to as *smoothing* a PV.

5.2.3 Advantages of smoothing predictors

To illustrate the advantage of modelling data using GAM as opposed to GLM, I present a comparison of models fit using a subset of the INFLAOS dataset containing only noun plural tokens. The DV is the orthographic realisation of the plural noun (npl) inflection, specifically, whether it is realised as a zero-morpheme as in the example in figure 5.3, or not.

FIGURE 5.3: Extract from text 1791 [1461, BUR, ABD].

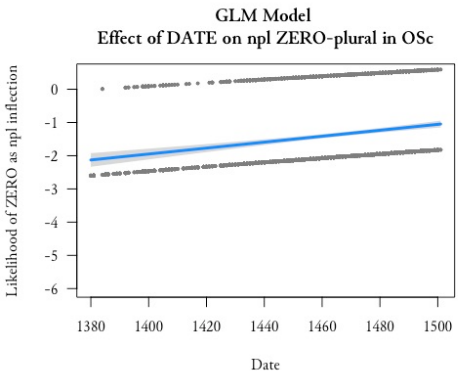


<to delyu(er) al thyng put to thai(m)>
to deliver all things put to them

As demonstrated in section 4.1.4, the values of this DV are binary, with a value of 1 indicating a zero-morpheme, and 0 indicating a non-zero-morpheme. I firstly model the occurrence of zero-morpheme {S} as a function of date, that is, I construct a model which reveals how well date (the PV) predicts the occurrence of zero-inflected npl {S} (the DV). The resultant plot of date against the log-likelihood of an npl token having a zero-inflection is shown in figure 5.4a.

FIGURE 5.4: Results of GLM and GAM models of the effect of date and etymology on npl zero inflections.

(A) GLM: univariate



(B) GAM: univariate

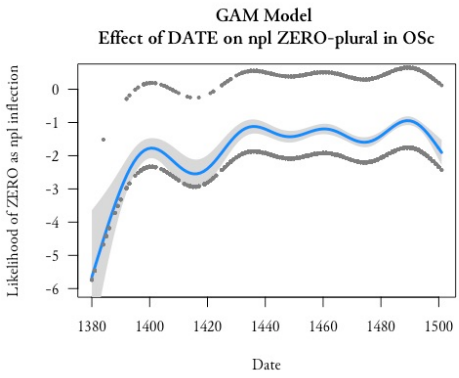


Figure 5.4a shows that the log-likelihood of an npl token having a zero inflection increases over time. Based on this graph, a token from a text written in 1380 is marginally less likely to have a zero-plural than a text written in 1500. However, a GLM is limited by the assumption that the relationship between the

dependent and PVs is linear. That is, the GLM models the effect of an increase in date on the likelihood of a zero-plural as a single coefficient value, attempting to assign the best fit based on the observed data. A GAM, on the other hand, does not rely on this assumption. Rather, a GAM models the dependent variable as a *smooth function* of the PV, allowing the effect of each *interval* of the continuous PV on the DV to be modelled. A GAM is fit to the same data as the GLM in figure 5.4a, and the resultant plot in figure 5.4b.[3]

It is clear that the plot resulting from the GAM leads to very different conclusions to that resulting from the GLM. Whilst the GLM plot suggests a steady increase in likelihood over the whole period, the GAM plot reveals that, whilst the likelihood does fluctuate throughout the entire period, the only overall increase is the steep increase at the beginning, from 1380 to around 1400. A GAM is able to model this kind of relationship by splitting the predicted effect of date on zero-plurals into intervals and estimating the relationship between predictor and dependent variable in each one. In figure 5.4b, the grey band around the blue regression line, which represents the 95% confidence interval for the model's prediction of the likelihood of a zero plural, is wider than at other points along the line, getting narrower as time goes on. A qualitative inspection of the data reveals that the distribution of tokens across time is skewed, with far fewer tokens in the first half of the fifteenth century than in the second (as shown in figure 4.6). This causes the confidence intervals to be wider at the points where tokens are scarce because there are fewer data on which the model can base its output for these periods.

The plot for the same data modelled using a GLM in figure 5.4a does not show a change in confidence intervals of this magnitude across the time period because it is only able to fit a linear regression line, as opposed as opposed to the smooth line fit by GAM. That is, the GLM uses the available data observations to 'decide' on a linear trend which best fits the data, whereas the GAM allows the data to 'speak for itself' by constructing a trend line based on smoothed-over functions which describe the relationship between the DV and the PV at different values of the PV. This is what is meant by the terms 'additive' and 'linear' models - in a linear model, the DV is assumed to be a linear combination of the PVs. In an additive model, this assumption is relaxed to allow the modelling of a non-linear relationship.

5.2.4 Multivariate analysis

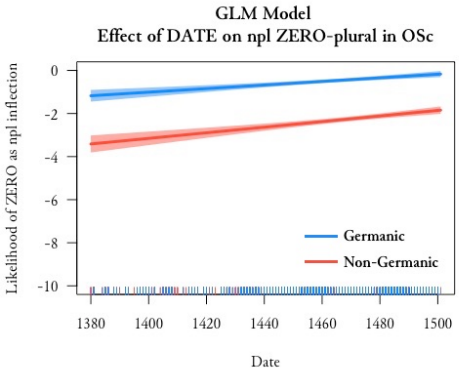
In the previous GLM and GAM examples, the dependent variable was modelled according to a single predictor, DATE. However, it is very often the case that there is more than one predictor of a particular phenomenon. To illustrate the effect this can have on the results of a regression analysis, another pair of GLM and GAM models is fitted, this time including an interaction between date and etymology. Specifying an interaction between two PVs means that the model will output estimates of: (a) the significance of the correlation of zero with date (controlling for the the correlation of zero with etymology and its interaction

[3]The R package used to fit GAMs throughout this investigation is mgcv (Wood 2006)

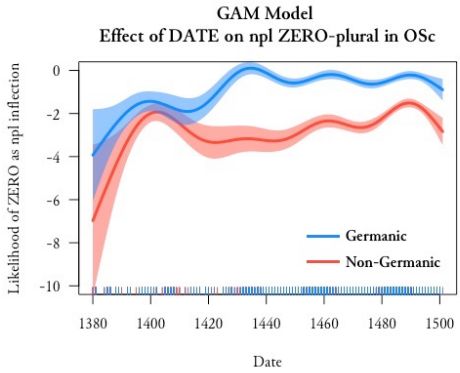
with date); (b) the significance of the correlation of zero with etymology (controlling for the correlation of zero with date and its interaction with etymology); and (c) the significance of the correlation of zero with date for each etymological category (Germanic and non-Germanic). Figure 5.5a and 5.5b show the plots resulting from the newly specified GLM and GAM models including the interaction of DATE with ETYMOLOGY.

FIGURE 5.5: Results of GLM and GAM models of the effect of date and etymology on npl zero inflections.

(A) GLM: multivariate



(B) GAM: multivariate



It is clear from both plots that overall, tokens of Germanic lexels are more likely to have a zero-plural inflection than tokens of Non-Germanic lexels. However, whilst the GLM plot in figure 5.5a appears to show a steady, overall increase in the likelihood for both categories, the GAM plot reveals that the likelihood of zero-plurals for Germanic and Non-Germanic lexels was very similar from the beginning of the period until around 1420, when the likelihood of zero-plurals increases for Germanic lexis but not for Non-Germanic lexis.

5.2.5 Mixed effects

In section 5.2.2, I introduced regression modelling and demonstrated the advantages of using GAM for an unbalanced time-series dataset such as INFLAOS, as opposed to using a linear modelling approach such as GLM. In this section, I present the advantages of using a mixed effects modelling structure. This is not something specific to GAM modelling, as it can be used in the context of linear modelling using generalised linear mixed effects model (GLMM).

However, as my analysis uses GAM, I restrict my explanation in this section to the inclusion of random effects in GAMs.

The term *mixed effects* refers to models which include both *fixed effect* PVs and *random effect* PVs. Fixed effect PVs are those which are included in the model in order to gauge their effect on the DV. The GAM

models in section 5.2.2 included only fixed effect predictors. These models attempted to return the best estimation of the effects of the predictors ETYMOLOGY and DATE based on the observed data. The *goodness of fit* of the model can be measured in terms of how much of the variation in the observed data was able to be captured by the predictors, with the remaining deviance attributed to random error in the model. Gries (2015: 97) succinctly describes the advantages of mixed effects models over fixed-effects-only models, stating that

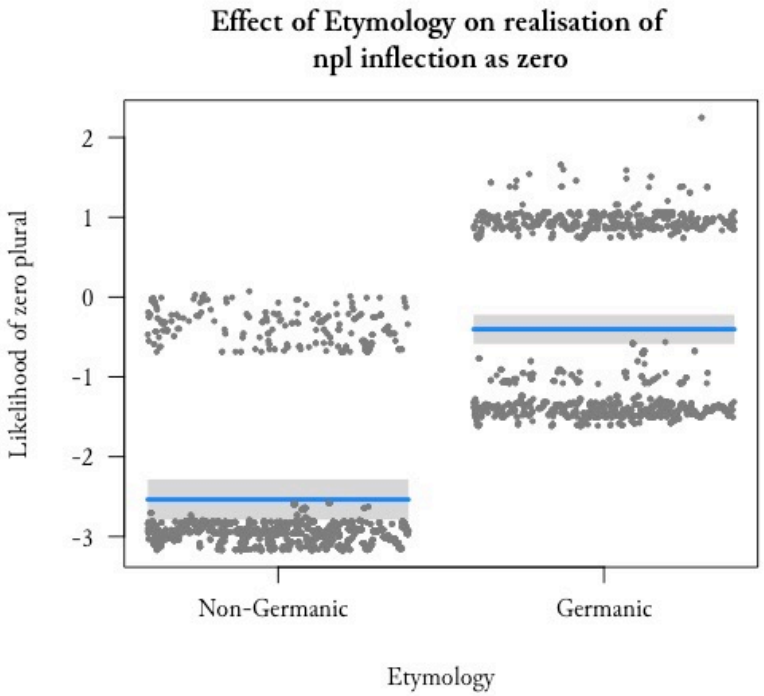
> [Mixed effects models] address the fact that data points are related because they were provided by the same subject or for the same item. [...] In other words, statistical analyses become more precise because [...] the idiosyncrasies of particular speakers [...] do not distort the regression coefficients, but are 'relegated to' the variance captured by what are called 'random effects'.

To relate this to the data analysed in section 5.2.2, consider the effect of ETYMOLOGY on occurrence of the zero plural as shown in figure 5.5b. The individual effect of ETYMOLOGY is shown in figure 5.6. The effect of etymology is shown by an Analysis of Variance (ANOVA) test to be significant at $p < 0.001$ (line 11). However, this model does not take into account any potential effect on the model prediction of individual lexels. Etymology appears to be a highly significant predictor of whether an npl inflection will be realised as zero or not, but the current model offers no way of ascertaining whether this apparent effect is genuinely due to a difference in inflection between Germanic and non-Germanic lexis, or whether the effect is in fact being produced by some particular words which fall into these categories.

As an exaggerated example, consider a dataset of OSc inflections *L*, including all of the same PVs as INFLAOS, but wherein 30% of the npl inflections are zero. However, all instances of this zero plural in *L* occur with the common Germanic lexel *land*. A model which does not account for similarity between tokens of individual lexels such as the one above would return a highly significant result for the effect of etymology on zero plurals given that zero plurals only occur with Germanic lexis. However, the model would not "see" the individual lexels, and therefore would not reveal that the dependent variable is predicted by a single lexel. This is an example of what Gries (2015: 97) refers to as "idiosyncracies", the unique behaviours of individual groupings of tokens (Gries refers to experimental responses grouped by speaker, but the idea is the same for tokens grouped by text or by lexel). In the case of this hypothetical dataset *L*, the variation in the data is due to the idiosyncratic behaviour of a single lexel, and is therefore not representative of all Germanic lexis.

The inclusion of random effects is crucial when modelling data which has a hierarchical structure (i.e. tokens grouped according to lexel, tokens grouped according to text) to ensure that significant fixed effect predictors remain significant when we account for the individual variation between tokens which are related to each other. The relevant relationships in INFLAOS are due to tokens coming from the same text, or

FIGURE 5.6: Result of GAM model: individual effect of etymology on npl zero inflections.

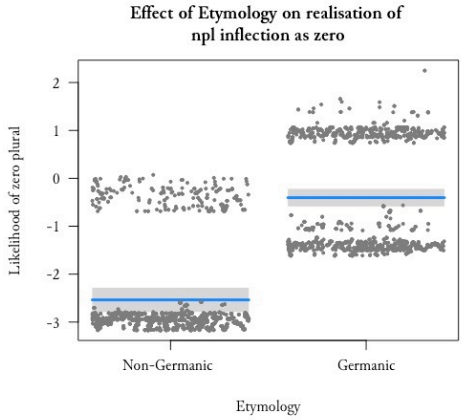


exemplifying the same lexel. A model of the effect of date and etymology on the occurrence of the npl zero plural, this time including the random effect PV lexel, is then specified. When the model including the random effect (gam.fit3) and the model excluding the random effect (gam.fit2) are compared using ANOVA, the difference between the models is estimated to be significant at $p<0.01$, indicating that the *difference between the two models* is highly significant. That is, the inclusion of the random effect of lexel in the model causes the model to fit the data significantly better than it did without the random effect. The p-values estimated for each of the model predictors by ANOVA indicate the significance of that predictor when all others in the model are held constant. Assuming a significance threshold of $p < 0.05$, it is easy to see that the only significant predictors in the model are now date (line 17) and the random effect of lexel (line 16). It would appear, then, that the significant effect of etymology indicated by the fixed-effects-only model was in fact due to individual lexical variation. Once this variation is accounted for systematically rather than anonymised by inclusion with the general random error of the model see section 5.2, the model no longer reports any significance of etymology as a PV. This difference in the predictive significance of etymology can be seen in a comparison of the graphical outputs of the fixed-effects-only and random-effects models.

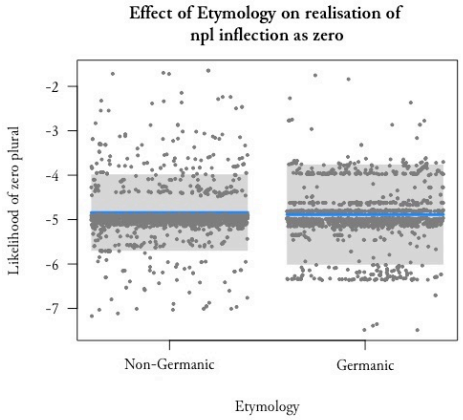
Figure 5.7a reproduces the graph of the individual effect of etymology from figure 5.6, whilst figure 5.7 shows the output of the model with the random effect of lexel included. When the random effect is not

FIGURE 5.7: Results of GAM models of the effect of date and etymology on npl zero inflections with and without the random effect of lexel.

(A) Fixed-effects-only model



(B) Fixed and random effects model



included, the likelihood of a zero plural occurring on a Germanic word is estimated to be significantly higher than on a non-Germanic word. When the random effect is included, the likelihood of a zero plural is estimated to be similar regardless of the etymology of the word.

5.3 Choice of method: GAM

This investigation makes extensive use of GAM modelling. As identified in section 5.2.2, the advantage of GAM as opposed to other regression modelling techniques is that the effect of PVs on the DV is not required to be linear. There are other techniques which make use of mixed effects, notably GLMM, which allows the hierarchical structure of PVs to be modelled as mixed effects. Using GLMM, however, it is only possible to estimate the linear effect of a PV on the DV. Recall the example using GLM in section 5.2.4 (figure 5.5a) to model the effect of date on npl zero inflection. A GLMM would have the advantage of allowing random effects to be included in the model. We might find, for example, that when the variation between individual texts is taken into account, the date of a text no longer appears as such a significant predictor. That is, the variation in the DV which seemed to correlate with text date could in fact be due to random variation between texts. However, a GLMM would still be limited to estimating a linear correlation between text date and the likelihood of npl zero inflection. In other words, there are three possible outcomes of estimating the effect of text date: (a) the likelihood of zero increases over time; (b) the likelihood of zero decreases over time; or (c) text date has no effect on the likelihood of zero. Conversely, using GAM to model the same data (as in figure 5.5b), it is possible not only to include random effects such as individual text, but also to estimate the correlation between text date and the likelihood of zero inflection at different

time periods. Figure 5.5b, for example, shows that the increase in likelihood of zero over time occurs largely before 1400, and that during the rest of the period to 1500, the likelihood of zero remained fairly stable. In the GLM of this data, the trend was shown as an overall increase throughout the period 1380 to 1500, which does not accurately reflect the trend shown in the data.

Part III

Results

Chapter 6

Irregular Inflections

6.1 Introduction

Though this investigation is concerned primarily with the orthographic realisation of the {S} and {D} inflections of plural noun (npl), present tense verb (vps), past tense verb (vpt) and past participle (vpp) forms, to fully understand their distribution, it is necessary to consider also the environments in which these morphemes are not realised at all. There are several categories of tokens which do not contain a realisation of the {S} or {D} morpheme. In section 6.2, I firstly discuss irregular forms of npl inflection, including weak {N} forms, mutative forms and {R} forms. These irregular forms make up over half of the non-{S} npl tokens in *Inflections in A Linguistic Atlas of Older Scots* (INFLAOS). The rest are zero-inflected tokens.

The distribution of zero npl, vps, vpp and vpp inflections in INFLAOS is shown in figure 6.1. Each column in this graph represents a grammel, and each word represents all the tokens of a single lexel tagged with that grammel in INFLAOS. The size of each word indicates the total tokens of that particular lexel-grammel combination in INFLAOS. The height of each word (its position on the y-axis) shows the percentage of tokens which have no {S} or {D} inflection realised in the manuscript. There is a clear difference between {S} and {D} observable in the graph: npl and vps lexels are spread out over a range of percentage values, whereas vpt and vpp lexels are not, instead forming two clusters at the extremes of the scale. This strongly suggests that the occurrence of zero in vpt and vpp tokens is conditioned by the lexels themselves. Specifically, lexels appear to, in general, take a zero inflection all of the time, or take it none of the time. This distinction corresponds to the regularity or irregularity of lexels. In other words, figure 6.1 shows that zero in vpt and vpp appears with irregular verbs, which always have zero, and not with regular verbs. There are few exceptions to this binary distribution. The plot shows only three lexels which do not form part of the

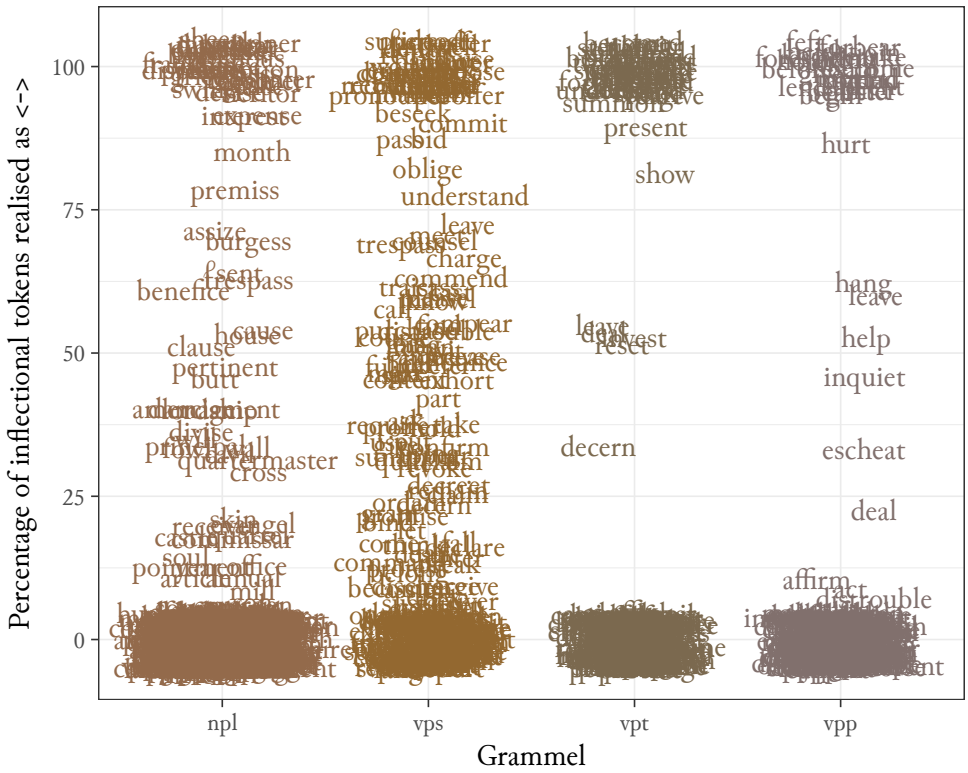
clusters at the extremes of the scale. These are:

(a) *show* (vpt) - mutative <schew> alongside zero <schow> and combination <schewit>, also <scwth>.

(b) *summon* (vpp) - Scots *summond*

(c) *leave* (vpp) - <left> also seemingly separable form <lefit> tagged in *A Linguistic Atlas of Older Scots* (LAOS) as LEF+IT, also forms suggesting stem-final voiced labio-dental fricative <levit> and <lewᵗ>.

FIGURE 6.1: Realisation of npl, vps, vpt and vpp inflections as zero. Each column reprsents a grammel and the position of each lexel on the y-axis indicates the percentage of tokens of that lexel with a zero-inflection. Only lexels with >10 tokens are included.



There are also clusters at the extremes of the npl column, though the cluster at 100% is smaller than those of vpt and vpp, indicating a smaller number of irregular npl lexels than vpt or vpp. It is also dominated in size by the lexels *man* and *witness*, which make up a large proportion of the irregular npl tokens in INFLAOS. There are many lexels occupying the area between the extremes of the scale for npl, indicating that they sometimes occur with zero and sometimes with a realisation of {S}. In contrast to the other three grammels, The distribution of vps lexels does not feature tight clusters at the scale extremities. Instead, the lexels are distributed more evenly on the scale, though more densely below 50% zero than above. This

distribution indicates that a great many vps lexels have zero inflections some, but not all, of the time. The only vps lexel which consistently occurs with a zero inflection is *will*. The fact that vps lexels are widely distributed in terms of percentage of zero is consistent with the operation of the Northern Subject Rule (NSR), which causes predictable and regular zero inflection.

Figure 6.1 shows that the occurrence of vpt and vpp zero is almost entirely orthogonal on the distinction between regular and irregular lexels. The dataset is close to perfectly separated, meaning that, with only a few exceptions, zero inflection is confined to irregular lexels and {D} to regular lexels. This chapter, therefore, focusses on npl and vps zero, with the aim of ascertaining the factors conditioning it.

Section 6.3 shows the percentage of npl and vps {S} tokens in each INFLAOS text and lexel which are realised as zero. There is more variation in the percentage of npl {S} tokens realised as zero across different lexels than across different texts, but this pattern is not replicated in the equivalent figures for zero realisation of vps {S}. Rather, the between-lexel and between-text variation in vps zero appears similar. These different inter-text and inter-lexel distributions suggest that the realisation of npl {S} as zero is predictable according to lexical predictor variables (PVs) whereas vps zero is not.

Section 6.3.1 describes the distribution of zero in npl tokens, beginning with the potential correlation of certain semantic categories of nouns suggested by Kopaczyk (as Bugaj 2002: 97). Whilst the specific lexels mentioned by Kopaczyk are found to correlate with zero in INFLAOS, the generalisation does not hold for other lexels which fall into the same categories. However, Kopaczyk's suggestion that the phonetic and orthographic similarity between the stem-final segments of particular lexels and fully-realised forms of {S} is supported by the INFLAOS data. This correlation of zero with final segments is further explored in the context of all stem-final *littera* (SFL), where stem-final <s> and <ß> are found to correlate with npl zero. The PVs of text date and type are also described in section 6.3.1. Text date does not show a correlation with npl zero, but there is a correlation between zero and texts identified as coming from notarial protocol books.

In section 6.3.2, a generalised additive model (GAM) is fitted incorporating the PVs described in section 6.3.1, as well as incorporating text location as a predictor by including a two-dimensional smooth function of latitude and longitude. Significant predictors of npl zero are found to be SFL, text type, date and location. The random effect of individual lexel is also found to be significant. The model summary is presented first, followed by a visualisation and qualitative analysis of the log-likelihood estimates for each of these PVs.

Section 6.3.3 begins by describing the distribution of vps zero inflection according to the grammatical contexts which predict the operation of the NSR: pronominality and adjacency of subject. Tokens with adjacent pronominal subjects are found to have a much higher percentage of zero inflection than other tokens. The correlation between SFL and percentage of vps {S} tokens realised as zero is also plotted.

Stem-final <e> and <ß> are more likely to be followed by zero than other SFL, including <s>, which indicates a difference between npl and vps in this regard.

The percentage of zero in vps tokens is also plotted according to text date and type. As with npl, no correlation is observable with date, but the higher percentage of npl zero in texts from notarial protocol books is not found for vps zero.

Section 6.3.2.1 presents the results of GAM model fit to the INFLAOS vps data. As expected based on the descriptive statistics, when all PVs are considered together, only NSR and SFL are significant predictors. The random effect of individual lexel is also found to be significant. The model summary is presented first, followed by a visualisation and qualitative analysis of the log-likelihood estimates for each of these PVs.

6.2 Irregular npl inflectional morphemes

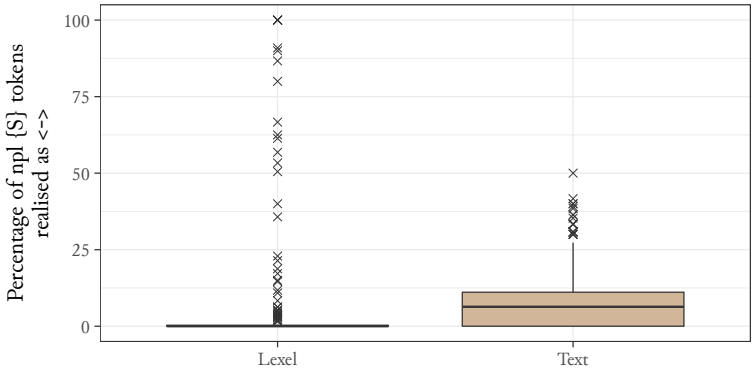
Irregular nouns account for a large majority of the npl tokens in INFLAOS which do not have an {S} inflection. There are 1,993 npl tokens in INFLAOS which do not feature an {S} inflection. Of these, *ox* > *oxen* (seven tokens) and *shoe* > *shoen* (one token) take an {N} inflection, a reflex of the Old English (OE) weak *-an* class (King 1997: 163). *Child* > *childer* (five tokens) takes an {R} inflection, a reflex of the OE *-ru* plural inflection, and contrasting with the English plural form *child* > *children*, which has both {R} and {N} inflections (King 1997: 164). A larger number of irregular npl tokens are those with mutative plural forms, including 579 tokens of *[wo]man* > *[wo]men*, or compounds headed by *-man*, such as *countryman* and *freeman*. There is one token of a compound headed by *-woman*, *gentlewoman*. Further mutative plurals are *brother* > *brether* (46 tokens), again contrasting with the corresponding English plural *brethren*, which exhibits both vowel mutation and a weak {N} inflection; *cow* > *kye* (11 tokens); *goose* > *geese* (two tokens); and *foot* > *feet* (one token).

6.3 Zero-inflected {S} tokens

Figure 6.2 shows the percentage of npl {S} tokens realised as zero in each lexel (left-hand plot) and in each text (right-hand plot). The mean percentage of tokens realised as zero is very low, whether it is represented by percentage per lexel or per text. However, there is more variation between individual lexels, as shown by the range of outliers in the lexel boxplot extending over the entire range of the percentage scale. In comparison, the points representing individual texts are much closer together, and group into a visible interquartile range. This difference suggests that the zero-realisation of npl {S} is predicted more accurately by lexical than by textual factors.

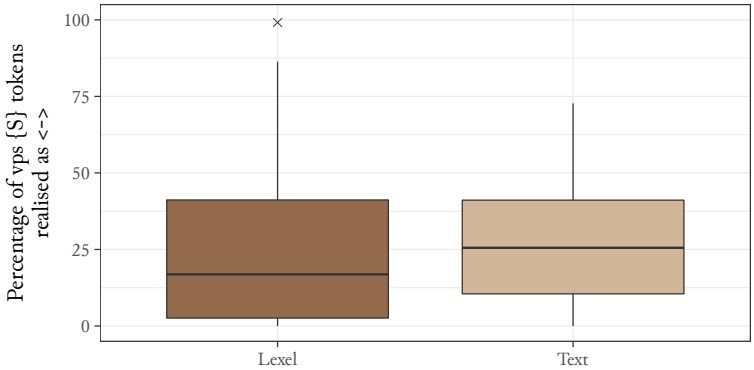
The equivalent plot for vps {S}, shown in figure 6.3, does not echo the distinction between text and lexel shown in figure 6.2. Instead, the distribution of points representing vps {S} zero for texts and lexels

FIGURE 6.2: Percentage of npl {S} tokens realised as zero. The left-hand plot is based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plot is based on the percentage by text (so individual data points represent individual texts). Only texts and lexels with more than 10 tokens are included.



is very similar.

FIGURE 6.3: Percentage of vps {S} tokens realised as zero. The left-hand plot is based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plot is based on the percentage by text (so individual data points represent individual texts). Only texts and lexels with more than 10 tokens are included.



Given that vps {S} tokens are subject to the operation of the NSR, the distinction between figures 6.2 and 6.3 makes sense. Because there is a grammatical rule governing the use of zero inflection for vps {S}, it is equally likely to be used in any text and with any lexel. Conversely, the realisation of npl {S} as zero is not the result of a productive grammatical rule, but rather the existence of specific lexels with irregular plural forms. The lexels which have these irregular plurals in many cases also have a regular form - this accounts for the outlying lexel points in figure 6.2 occupying different positions on the zero-inflection percentage scale. Whilst irregularity is a lexical PV, it is likely that there is an element of scribal choice in the use of zero plurals, as suggested by the presence of text outliers in figure 6.2.

6.3.1 Zero-inflected npl tokens

Excluding the weak, mutative and {R} plurals described in section 6.2 leaves 1,307 remaining npl tokens which do not have an {S} inflection. These are zero-inflected tokens. King (1997: 163) points out that Older Scots (OSc) had more irregular, zero plural nouns than Modern Scots (ModSc) now has, and that these nouns were often those descended from strong, neuter OE nouns, which had a zero plural form. Further to this generalisation, Kopaczyk (as Bugaj (2002: 97)) identifies six semantic groups which characterise the zero inflections in the early sixteenth century OSc of the *Wigtown Burgh Court Book*. These categories are listed in table 6.1 along with the lexels attested in INFLAOS which denote nouns specifically placed in these categories by Kopaczyk.

TABLE 6.1: Lexels attested in INFLAOS which fit into one of the six semantic categories of zero npl tokens proposed by Kopaczyk (as Bugaj (2002: 97))

Semantic group	Lexels	INFLAOS zero	INFLAOS IS	Total
animals	*sheep*	14	0	14
	swine	13	0	13
	horse	11	0	11
	deer	1	0	1
	fish	6	0	6
	fowl	1	2	3
time span	*year*	137	484	621
	month	13	2	15
	term	2	337	339
groups of people	*witness*	465	0	465
	burgess	17	8	25
space measurement	e.g. *acre*	(all examples tagged as nqpl in LAOS)		
quantity or weight	e.g. *crown*	(all examples tagged as nqpl in LAOS)		
units of measurement	e.g. *firlot*	(all examples tagged as nqpl in LAOS)		
number		(all examples tagged as qc (cardinal quantifier) in LAOS)		
Total		**680**	**833**	**1,513**

The most striking observation from this table is that 465 zero plurals are tokens of the lexel *witness*. Kopaczyk (as Bugaj (2002: 102)) characterises the semantic group to which *witness* belongs as "groups of people", a category containing only one other noun, *burgess*. However, the reason suggested by Kopaczyk for the unmarked plural forms of these words is the fact that they end in *-es*, which "could have appeared awkward to the scribe when a regular plural suffix was added". This suggestion appears to be corroborated by the INFLAOS data. Table 6.2 lists the lexels attested with npl forms in INFLAOS which have a stem-final *litteral* string resembling the regular {S} inflection form.

83% of these forms have zero inflections. If *witness* is included, this figure increases to 98%. For npl forms in INFLAOS overall, the level of zero inflection is approximately 10%. If this constraint is affecting the realisation of the plural inflection then, according to the INFLAOS data in table 6.2, it is doing so

TABLE 6.2: Lexels with stem-final *litteral* strings similar in sound and orthographic realisation to attested realisations of npl {S}.

Lexel	Stem-final strings	Zero tokens	Total tokens
absent	<eß> (1), <esß> (1)	2 (100%)	2
assize	<iß> (2), <is> (1)	2 (67%)	3
burgess	<es> (10), <esß> (3), <is> (1)	14 (100%)	14
case	<iß> (1)	1 (100%)	1
commissar	<isß> (10)	10 (100%)	10
damage	<is> (4), <yß> (1)	5 (100%)	5
devise	<iß> (1)	1 (100%)	1
divise	<is> (4), <iß> (3), <ys> (1)	3 (38%)	8
fortalice	<es> (2)	0	2
fortress	<ess> (1)	0	1
franchise	<iß> (1)	1 (100%)	1
grice	<iß> (1)	1 (100%)	1
interest	<esß> (10)	10 (100%)	10
mass	<ess> (4)	0	4
pertinent	<eß> (1)	1 (100%)	1
premiss	<isß> (7), <iß> (1), <is> (1), <iss> (1)	8 (80%)	10
promise	<iss> (1)	0	1
witness	<es> (308), <eß> (125), <esß> (9), <is> (2), <iß> (1)	445 (100%)	445
Total			**520**

independently of the semantic categorisation of "groups of people" proposed by Kopaczyk, as the other lexels which constitute the evidence for this phenomenon do not fit into this class.

The semantic categories proposed by Kopaczyk account for a total of 680 zero npl forms in INFLAOS, a figure largely reflecting the high frequency of zero-inflected tokens of *witness* and *year*. If the final category of "groups of people" is reanalysed as scribal aversion to repetition of <es> or <is> strings, as is suggested by table 6.2, a further 57 tokens are accounted for.

Figure 6.4 shows the percentage of npl tokens of each INFLAOS lexel with a zero inflection. The lexels are grouped according to the categories proposed by Kopaczyk. The box plot for the category 'animal' shows that all tokens of lexels denoting animals are realised with a zero morpheme. However, the plot only shows points corresponding to lexels with more than 10 tokens. The only lexels represented are therefore *horse*, *sheep* and *swine*. Table 6.3 shows the full picture of animal lexels in INFLAOS, the majority of which have only one or two tokens. Excluding the animal lexels mentioned by Kopaczyk, very few have zero.

The next box plot in figure 6.4 groups the lexels which denote people. The distribution of lexels in this plot clearly supports Kopaczyk's finding that *witness* and *burgess* take a zero inflection, as well as her suggestion that this fact may have more to do with the phonetic similarity of the stem-final string to {S} than the actual semantic grouping. INFLAOS has many more lexels denoting people than were available to Kopaczyk in the *Wigtownshire Burgh Court Book*, and is therefore better able to show the contrast between

FIGURE 6.4: The percentage of npl tokens of each INFLAOS lexel with a zero inflection. The lexels are grouped according to the categories proposed by Kopaczyk (as Bugaj (2002: 97)).

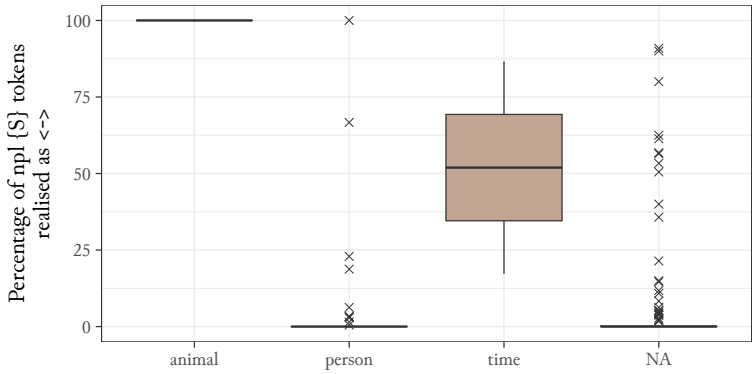


TABLE 6.3: Animal lexels in INFLAOS with token counts.

Non-zero				**Zero**	
calf	3	goat	1	sheep	14
mart	3	haddock	1	swine	13
cattle	2	lamb	1	horse	11
chicken	2	ram	1	fish	6
ewe	2	salmon	1	deer	1
fowl	2	spelding	1	fowl	1
hog	2	trout	1	grice	1
boar	1	veal	1	grilse	1
cod	1				
Total non-zero			**26**	**Total zero**	**48**

these two lexels and others denoting people.

The third plot in figure 6.4 groups lexels denoting time. As with animal lexels, however, lexels denoting periods of time are scarce in INFLAOS. The plot contains only two points, the lexels *year* and *month*. *Year* has many more tokens than *month* - 797 for *year* compared with 15 for *month*. The range of the plot corresponds to the difference between the two lexels in terms of percentage of zero tokens, with *month* at the higher extreme: 13 tokens out of 15 have zero. *Year* is at the lower extreme of the plot, with 137 tokens with zero of a total 797. The only lexel which has been omitted from the 'time' category due to low token frequency is *hour*, which has only three tokens, none of which are zero.

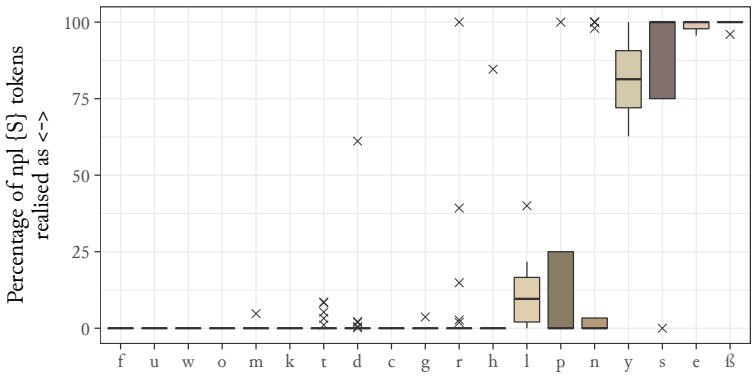
The final box plot in figure 6.4 represents the percentage of zero tokens attested for lexels which do not fit into the categories 'animal', 'person' or 'time'. The mean percentage value is zero, with various lexels appearing as outliers at different percentage values. In the interest of clarity, only those outlying lexels with percentage values above 25% are labelled. Some of the outliers also appeared in table 6.2 which listed all forms ending in a *litteral* string resembling {S}: *interest* (tokens tagged with this lexel are examples of

the earlier form *interes(se)*); *premiss*; and *damage* (this lexel is often used to tag examples of OSc *dampnis* 'damages', a static form which always had a plural referent (*Dampnis n. pl.* 2004)). The lexels *absent* and *pertinent* are also attested with a fairly large percentage of zero tokens. In total, 102 out of 202 tokens of plural *pertinent* have a zero ending. The majority of these zero-inflected tokens are of the form <p(er)tinen> with a stem-final horizontal siglum. The *Dictionary of the Scots Language* (DSL) contains several examples matching this n-final form, but all are listed as examples of the word *pertinence* (*Pertin-, Pertenence, -ens n.* 2004), and only forms with stem-final <t> are listed under *pertinent* (*Pertinent n.* 2004). Contextually plural instances of <ns>-final forms are suggested to be colloquial, uninflected forms of *pertinence*. An examination of the individual LAOS texts where this form of plural *pertinent* appears suggests that this form may indeed be idiosyncratic - 75 of the 89 <ns>-final forms are from notarial protocol books which are attributed to the same author - Sir Thomas Crawfurd.

There are also lexels which are phonetically [ɪs]-final but are generally written with a stem-final <c> in INFLAOS, such as *benefice* and *office*. There are also several outliers ending in <l>/[l]: Germanic *will*, *mill* and *soul*; and non-Germanic *evangel*, *quarrel*, *castle*, *article*, *annual* and *seal*.

Figure 6.5 shows a series of box plots which represent the percentage of tokens realised with a zero inflection, grouped according to their SFL. There is a clear dichotomous distribution shown by the boxplots: the mean percentage value of all SFL except <y, e, s> and <ß> is zero, whereas that of <e, s> and <ß>-final tokens is very close to 100%. The high percentage of zero for <s> and <ß> is in line with the *horreur aequi* hypothesis proposed by Kopaczyk (as Bugaj 2002, but the interquartile range of 25% for <s> as opposed to 0% for <ß> suggests a difference between the two, namely that <ß> almost categorically occurs with zero, whereas <s> is very likely to be followed by zero, but this is not the case all of the time. The high percentage of zero for SFL <e> is due to its representing only two lexels, *benefice* and *year*.

FIGURE 6.5: The percentage of tokens realised with a zero inflection, grouped according to stem-final *littera*.



The only two SFL which do not conform to this extreme distribution are <l> and <y>. All data for <y>-final tokens are grouped at approximately 60%, and <l> data have a mean of approximately 10%, with

more variation between individual lexels suggested by the interquartile range of approximately 15%. *Will* is an outlier for <l>, with approximately 40% of *will* tokens realised with zero. The fact that *will* is often realised with a double final <l>, <ll>, provides a clue to the reason for this. It is noted by Williamson (2008: within LAOS text 835) that stem-final <ll> could be written with a cross-stroke through both letters to indicate "abbreviation of inflexion". On closer inspection of the individual data tokens, we find that there are four tokens of *will* with zero in INFLAOS and six with a realised inflection. The four instances with zero all have a stem-final horizontal siglum which may in this instance indicate plurality. Looking at the individual <y>-final tokens reveals that this lack of variance around the mean is due to there being only a single lexel representing SFL <y>, *other*. There are 260 tokens of *other* with stem-final <y>, all of which have a following siglum <ˈ> indicating the omission of <er> (Johnson and Jenkinson 1915: 59).

Because the same lexel can potentially have various orthographic realisations ending in different *litterae*, there are instances where the same lexel appears twice in figure 6.5 as part of the plot of two different SFL. For example, *other* is shown as an outlier in the plots of both SFL <d> and <r>. Examples (17) and (18) show instances of *other* with a final <d> and <r> respectively. The <d>-final realisation has a siglum following the SFL which represents <er>. These two appearances of *other* in figure 6.5 are close to one another on the y-axis, indicating a small degree of difference between the percentage of <d>-final and <r>-final tokens of *other* with zero.

(17) <Jhoñwyllsoñ alysond(er) bell vy^{t} vd(er) may & sy(n)dry>
Jhon Wyllson [and] Alysonder Bell, with others many and sundry. [text 1296: 1463, BUR, FIF]

(18) <ilkane til othyr haff gyfynē ye bodely aith̄>
Each one to [the] others have given the bodily oath [text 99: 1490, CHA, PTH]

Some other lexels which appear twice in the plot are very different with regard to percentage of zero:

(a) *cause* is an outlier in both the SFL groups <s> and <ß>, but whereas <s>-final lexels have a mean zero token percentage of close to 100% and *cause* is an outlier at 0%, tokens of *cause* with stem-final <ß> still have a percentage value of close to 100%. There are 52 tokens of *cause* in INFLAOS, of which 30 have stem-final <ß> and 22 have stem-final <s>. All but one of the <ß>-final instances has no following inflection, whereas all 22 <s>-final instances do have a following inflection. For tokens of *cause*, then, there is a clear difference between the realisation of npl inflection where the SFL is <s> and where it is <ß>.

(b) *pertinence* and *absence* are both outliers for SFL <t> and <n>.

Figure 6.6 shows the percentage of npl tokens in each INFLAOS text with a zero inflection in each year covered by INFLAOS, 1380-1500. Each point represents a text, and only texts with more than 10 npl

tokens are included to avoid extreme percentage values caused by texts with a very low number of tokens. As well as the individual text points, the graph shows two regression lines fit to the data. The red line fits the most accurate linear trend line to the data, and the blue line fits the most accurate smooth trend line; that is, a line which accounts for different correlations between date and percentage of zero per text at different times throughout the period. It can be observed that these two lines follow one another closely, suggesting that the relationship between date and percentage of zero is linear. However, though the linear trend line appears to show a slight positive overall correlation between date and zero, the smooth trend line diverges from it in both directions at various date points, particularly in the last 25 years. The positive correlation may therefore be insignificant.

FIGURE 6.6: The percentage of npl {S} tokens of each lexel in INFLAOS realised as zero between 1380 and 1500. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Red: linear trend line; blue: smooth trend line.

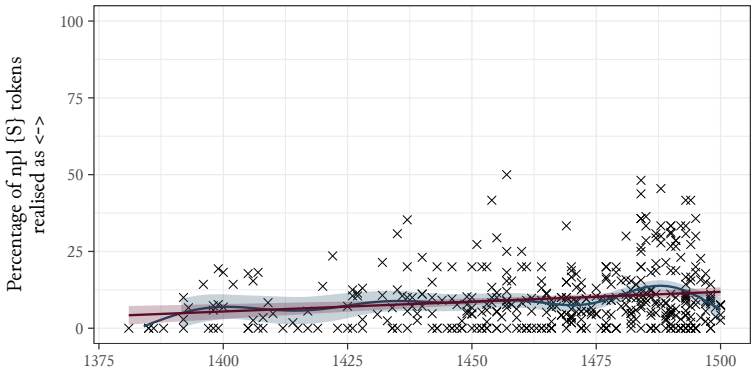
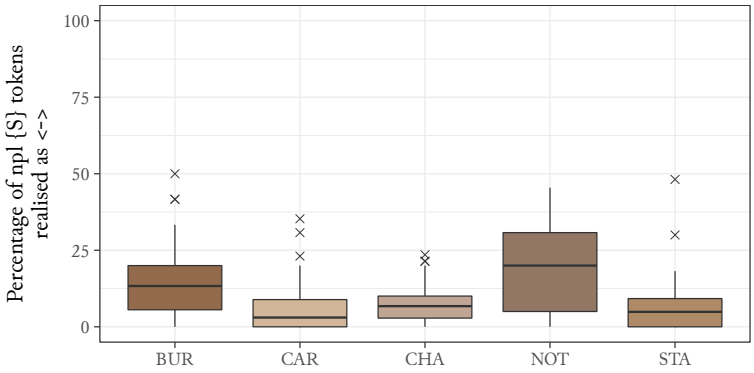


Figure 6.7 also shows the percentage of npl tokens in each INFLAOS text with a zero inflection, this time grouped by text type. Each box plot represents a text type, with outlying points labelled according to the text they represent. The only noticeable trend in the shapes of the box plots is that the median of the notarial protocol book (NOT) plot is higher than that of other types. This indicates a higher overall percentage of zero in notarial protocol books than in other types of text. Having said that, an examination of the outlying text points of the other type plots suggests a reason for this. The outlying texts are those which happen to have an unusually large number of plural tokens of one of the lexels *witness* or *year*. For example, the average number of tokens of plural *witness* in burgh records (BUR) is 1.3, but the outlying Text 1840 contains 10 such tokens which account for all of the zero npl tokens in the text.

Table 6.4 shows the number of texts of each type in INFLAOS and the number of those texts which contain at least one token of plural *witness*. Notarial protocol books have the highest percentage of texts containing plural *witness* (47%). Charters (CHA) also have a high percentage of texts containing *witness* (42%), however their lower average can be attributed to the fact that charters have, on average, one-third more words per text than notarial protocol books (see section 4.1.5). Table 6.4 also shows the average

FIGURE 6.7: Realisation of npl inflections as zero. Only lexels with >10 tokens included.



frequency of plural *witness* tokens for each text type, based only on texts which have at least one token of plural *witness*. The average for most of the text types is close to the overall average figure of 1.2 tokens per text, indicating that texts typically have only one or sometimes two tokens of plural *witness*. The average for state documents (2.7) is higher due to the tokens occurring in a small number of texts, some of which have a high count of plural *witness* tokens.

TABLE 6.4: Tokens of plural *witness* in INFLAOS.

Text type	**Total texts**	**Texts with *witnesses***	**Avg. tokens of *witnesses***
BUR	624	162 (26%)	1.3
CAR	110	22 (20%)	1.2
CHA	301	125 (41%)	1.1
NOT	110	52 (47%)	1.1
STA	56	9 (16%)	2.7
Total	1,201	370 (31%)	1.2

6.3.2 Modelling the likelihood of zero realisation of npl {S}

Model 1 is a generalised additive model of zero inflection in INFLAOS npl tokens. The PVs included in the summary table are those which were found to be significant using a backwards-elimination method of model checking. That is, an initial model was constructed containing all of the PVs; then a model with the least significant PV removed was fitted. These two models were compared using Analysis of Variance (ANOVA) to obtain a p-value indicating the statistical significance of removing the PV in question. A p-value from the ANOVA of >0.01 indicated that removing the PV from the model did not make the model a significantly worse fit to the data. In other words, the PV did not add a significant amount of explanatory power to the model to make it worth retaining. This elimination procedure was repeated until all the PVs in the model were significant at p<0.01.

Model 1: The results of a generalised additive model of zero inflection in INFLAOS npl tokens.
R^2 = 0.85; Deviance explained = 84.4%; N = 11,896

Parametric terms

		Estimate	SE	z	
(Intercept)		-7.03	0.86	-8.13	***
SFL	c	-0.26	1.24	-0.21	
	e	4.75	1.04	4.56	***
	g	0.85	1.19	0.71	
	h	1.9	1.29	1.47	
	k	0.91	1.64	0.56	
	l	3.5	0.98	3.55	***
	m	1.82	1.21	1.5	
	n	3.15	0.93	3.4	***
	o	0.53	1.53	0.35	
	p	1.28	1.53	0.84	
	r	0.58	0.99	0.59	
	s	5.37	1.05	5.13	***
	ß	14.6	1.76	8.32	***
	t	-1.6	0.99	-1.62	
	y	1.82	1.77	1.03	
Type	BUR	0.5	0.22	2.32	*
	CAR	0	0.27	-0.02	
	NOT	1.37	0.33	4.12	***
	STA	0.5	0.35	1.44	

Approximate significance of smooth terms:

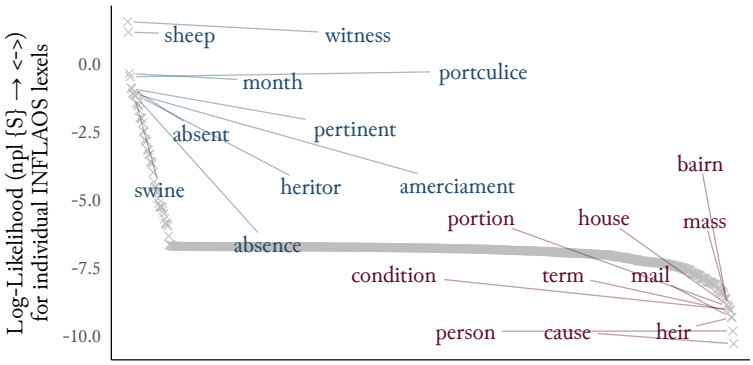
	edf	Ref.df	χ^2	
Lexel	141.68	776	639.2	***
Date	4.09	9	380.3	***
Latitude, Longitude	1.51	29	14.3	**

6.3.2.1 Individual lexels

Figure 6.8 shows the estimated log-likelihood of individual lexels to have a zero npl inflection. Lexels which appear higher on the y-axis are estimated by the GAM to have an unusually high log-likelihood of zero considering all other significant factors in the model. The plot therefore visualises the random variation in the GAM caused by individual lexels. The top and bottom 10 lexels are labelled.

Among the top 10 highest outliers in figure 6.8 are some nouns which took a zero plural form as standard in OSc, as discussed previously. Of these, *sheep* and *swine* retain a zero plural form in ModSc and Present Day English (PDE), and *month* retains a zero plural form in ModSc and some dialects of PDE (*Month n.1* 2004). Also among the top 10 are *absent* and *pertinent* which, as discussed previously, have forms which are ambiguous with *absence* and *pertinence* respectively. The lexel *absence* also appears as an outlier, though there is only one plural token of this lexel in INFLAOS. Lexels which appear as outliers but have not been mentioned previously are *heritor* and *amerciament*. A qualitative analysis of the data shows that *heritor* is found only once in INFLAOS. *Amerciament* is attested five times, and the likely reason for

FIGURE 6.8: Fitted values for the random effect of individual lexel in a generalised additive model showing the significant effects of etymology, SFL and NSR conditions on the realisation of npl morphemes as zero in INFLAOS.



its appearance as an outlier is the extremely truncated abbreviated form <aṁ>, used by the scribe or scribes of two texts, T1823 [BUR, 1489, FIF] and T1829 [BUR, 1489, FIF], both from the *Burgh Court Book of Dunfermline* and dated February and April 1489 respectively.

Apart from *witness*, only one of the annotated outliers typically ends in a *litteral* string resembling {S}: *portculice*. This may seem surprising in light of the conclusions previously drawn from table 6.2, which suggests that stems ending in an {S} -like string are more likely to be followed by a zero plural. However, it can be observed that the upper left tail of the plot in figure 6.8 is extremely long and steep, indicating the presence of many more lexels which are attested with zero plural forms more often than the model would suggest. The fact that the points in the plot form a fairly uniform line before suddenly diverging suggests that the likelihood of abbreviation is generally very predictable, but with certain lexels representing significant outliers. This characterisation fits with the general conclusions gleaned from the summary of Model 1, that zero realisation of npl {S} is highly predictable based on the fixed-effect PVs included in the model, and that the individual variation between lexels explains a significant amount of the deviance in the INFLAOS npl data.

6.3.2.2 SFL

The final model reveals that the PVs which significantly correlate with the likelihood of zero are SFL and text type. SFL has 19 categories in total, of which six are estimated by the model to be significant. In this context, an SFL being described as *significant* means that the likelihood of a token ending in that SFL being realised as zero is significantly different from the likelihood of a token ending in the reference category SFL being realised as zero. In this model, the reference category (see section 5.2.1) is set as <d>. The coefficient value for each SFL indicates the difference in log-likelihood of zero between tokens ending in that SFL and tokens ending in <d>. Each coefficient value has a corresponding standard error value, indicating the standard deviation within the category. In other words, how much the individual coefficient estimates of

tokens with that SFL differ from the mean on average. A lower value means we can have greater confidence in the coefficient estimate for that SFL, because there is less variation between individual data points - they behave more like a coherent category. The model also provides a *z* for each coefficient, indicating the number of standard deviations from the population mean. This value is used to judge the significance of a coefficient - whether it is significantly different from the mean. In the summary table for Model 1, *z* valuess with stars following them represent coefficients which are significantly different from the mean at p<0.01.

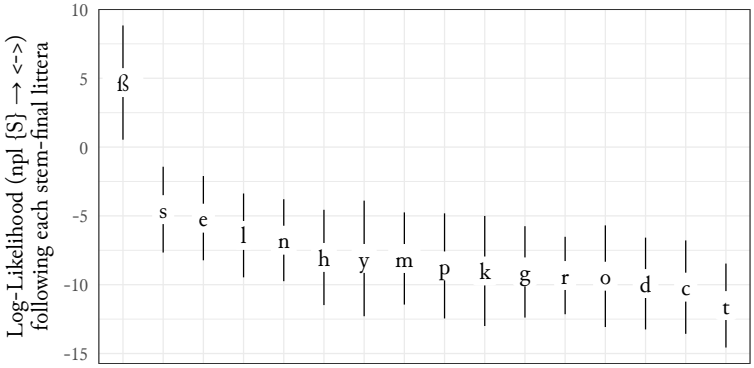
For example, <l> has a coefficient value of 3.52, indicating the difference between the log-likelihood of <d> being followed by zero and <l> being followed by zero. The standard error (or standard deviation) is 0.99, meaning that on average, individual estimates of likelihood of zero for <l> tokens vary from one another by 0.99. The *z* for <l> is 3.57. This means that the coefficient value for <l> (3.52) differs from the mean of the coefficients of other categories by 3.57 standard deviations. This *z* is significant at p<0.01. On the other hand, <r> has a coefficient value of 0.59 and a standard error of 0.99, meaning that the difference between the log-likelihood of <d> being followed by zero and <r> being followed by zero is 0.59, but the individual coefficient estimates for <r> tokens, on average, vary from each other by 0.99. The *z* tells us that the coefficient value for <r> differs from the mean of the coefficients of other categories by 0.596 standard deviations, and is not significant at p<0.01.

Figure 6.9 shows the log-likelihood estimates for each SFL presented on a graph. This allows us to examine the relative likelihood of zero following each SFL visually rather than as a value representing difference from the reference category. The position of each SFL on the y-axis represents the estimated log-likelihood of a stem ending in that SFL being followed by zero, and the vertical lines extending from each SFL point indicate 95% confidence intervals for each SFL's estimated log-likelihood. Some SFL have been omitted from figure 6.9: <f, u, v> and <w>. These SFL have extremely low coefficient values relative to the other SFL, and extremely high standard error values. This is due to these categories of SFL being *perfectly separated*, that is, all tokens ending in these SFL have the same value of the dependent variable (DV). In this case, the value is 'not zero'. This separation prevents the model from calculating an appropriate likelihood estimate and results in inflated estimates like those shown in the summary table of Model 1.

The most noticeable feature of figure 6.9 is that <ß> has a higher log-likelihood of being followed by zero than any other SFL. The SFL with the next highest log-likelihood value is <s>, though it is still far more in line with the other SFL categories than with <ß>.

Comparing the correlation between SFL and npl abbreviation shown by the descriptive plot in figure 6.5 with the results of the GAM, it is clear that the apparent effect of SFL is largely accounted for by the variation between individual lexels. The exception to this generalisation is the effect of stem-final double-s: <ß>. The descriptive plot in figure 6.5 appeared to show a similar positive correlation of zero-inflection with stem-final <s> and with stem-final <ß>. However, modelling the likelihood of npl tokens being realised

FIGURE 6.9: Estimated log-likelihood of zero for tokens ending in each stem-final *littera* (SFL).



with a zero inflection whilst controlling for between-lexel variation reveals that stem-final <ß> is a much stronger predictor of zero-inflection than stem-final <s>.

Table 6.5 shows the distribution of orthographic forms of npl {S} following stem-final <ß> compared to following other SFL. Stem-final <ß> is proceeded by a zero inflection in all but one instance. The table also shows the distribution of {S} forms following stem-final <s>. Modelling the likelihood of zero inflection in the data has already shown that the difference in percentage of zero inflections following stem-final <s> and <ß> is significant. Table 6.5 relates this finding to the wider context of possible realisations of npl {S}, and shows that not only is stem-final <ß> a significant predictor of zero inflection, but it almost exclusively occurs with zero. The fact that both stem-final <ß> and <s> are likely to be followed by a zero inflection, but <s>, unlike <ß>, also occurs in the data with fully-realised inflection forms.

TABLE 6.5: The distribution of orthographic forms of npl {S} following stem-final <ß> compared to following other SFL.

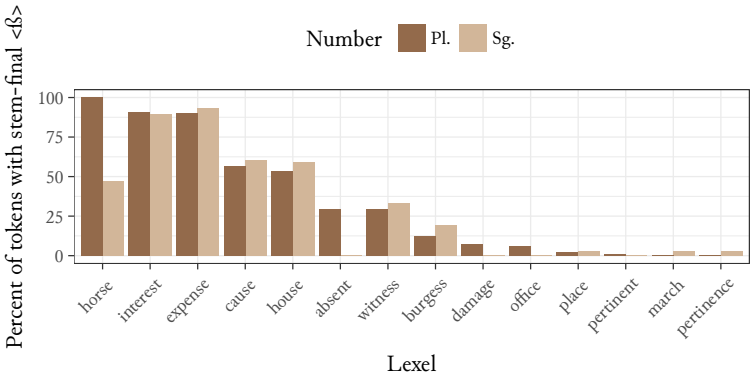
SFL	ß	s	Other	Total
<->	346 (31.10%)	355 (31.92%)	411 (36.96%)	1,112 (100%)
<ʃ>	0	3 (0.04%)	7,498 (99.96%)	7,501 (100%)
<C>	0	3 (0.56%)	536 (99.44%)	539 (100%)
<iC>	1 (0.04%)	50 (1.80%)	2,725 (98.16%)	2,776 (100%)
<yC>	0	0	577 (100%)	577 (100%)
<eC>	0	11 (1.60%)	677 (98.40%)	688 (100%)
Total	347 (2.60%)	422 (3.20%)	12,424 (94.17%)	13,193 (100%)

Given this difference in the inflections following stem-final <s> and <ß>, and the fact that a doubled <s> bears notional similarity to a stem-final single <s> with an {S} inflection, it seems possible that stem-final <ß> is functioning as an inflectional form in itself. To test this hypothesis, I analysed singular noun tokens extracted from LAOS to discover whether the occurrence of stem-final <ß> in singular tokens matched that in plural tokens. I hypothesised that if the lack of overt plural marking following stem-final <ß> is due to the *littera* itself performing that function, in the same way as other *littera* are commonly suspended or truncated

in OSc texts, then the occurrence of stem-final <ß> in grammatically singular nouns would be unlikely to follow the same pattern. On the other hand, if the absence of non-zero inflections following stem-final <ß> is due to scribal aversion to placing an {S} inflection immediately after <ß>, then the occurrence of stem-final <ß> in singular and plural nouns should be similar.

Figure 6.10 compares the percentage of <ß>-final singular and zero-inflected plural tokens of a subgroup of noun lexels. The subgroup contains lexels which: (a) have more than 10 singular tokens attested in LAOS; (b) have more than 10 plural tokens attested in LAOS; and (c) have at least one token with stem-final <ß> in either the singular or the plural.

FIGURE 6.10: The percentage of <ß>-final singular and zero-inflected plural tokens of a subgroup of noun lexels.



For each lexel, figure 6.10 shows the percentage of singular tokens with stem-final <ß> and the percentage of zero-inflected plural tokens with stem-final <ß>. In the majority of cases, these two values are remarkably similar. The only significant variation between singular and plural attestation of stem-final <ß> is with the lexels *horse* and *absent*. *Horse* has stem-final <ß> in 100% of its plural occurrences, but only in around 50% of its singular occurrences. However, this may be due to the fact that there are only 11 plural occurrences of *horse* in LAOS, but 32 singular occurrences. Similarly, *absent*, which has 0% stem-final <ß> in the singular but around 30% in the plural, has only 12 singular occurrences compared to 54 plural occurrences. These 12 singular occurrences all come from the same manuscript, the *Aberdeen Council Register* and the same two-year period, 1456-1457. The percentage of stem-final <ß> in this lexel, therefore, could very well be representative only of a single scribe's practice.

In general, it appears that where stem-final <ß> occurs in zero-inflected plural nouns, it is reflective of the same phenomenon in the corresponding singular noun. The high likelihood of zero-inflection following <ß> shown in figure 6.9, therefore, suggests that the zero-realisation of {S} is predicted by the usual occurrence of stem-final <ß> in the singular noun, and is not a representation of {S} in itself.

6.3.2.3 Text type

Figure 6.11 shows the log-likelihood of zero realisation of npl {S} tokens according to text type. As indicated by the summary statistics for Model 1, the only category of text which stands out is notarial protocol book (NOT). Notarial protocol books appear more likely than other texts to contain zero realisations of npl {S}.

FIGURE 6.11: The log-likelihood of npl {S} being realised as <-> for each text type, estimated by Model 1.

Table 6.6 shows the percentage of npl {S} tokens in notarial protocol book texts which are realised as zero, grouped according to the specific notary public (see chapter 4; table 4.12) who is identified as the author of the text in the Index of Sources (IoS). There are three notaries public named in the IoS: Sir Thomas Crawfurd; Peter Marche; and James Young. All 106 notarial protocol books transcribed in LAOS are attributed to one of them, though Crawfurd and Young account for the majority of texts (43 and 51 respectively), with only 12 texts attributed to Young. Based on the information in table 6.6, it appears that the correlation between notarial protocol books and likelihood of npl {S} zero occurrence can be attributed mainly to one notary: Sir Thomas Crawfurd, as 23% of his npl {S} tokens are realised as zero compared with Peter Marche's 7% and James Young's 5%.

TABLE 6.6: the percentage of npl {S} tokens in notarial protocol book texts which are realised as zero, grouped according to notary public.

Notary	Texts	Non-zero tokens	Zero tokens	Total tokens
Crawfurd	43	428 (77%)	130 (23%)	558 (100%)
Marche	12	153 (93%)	11 (7%)	164 (100%)
Young	51	441 (95%)	23 (5%)	464 (100%)
Total	**106**	**1,022 (86%)**	**164 (14%)**	**1,186 (100%)**

At face value, these figures seem to suggest that Crawfurd is a 'heavier user' of zero npl {S} than the other two notaries. However, an examination of the specific npl lexels used by each notary reveals a different pattern, which is illustrated in figure 6.12. Figure 6.12 shows the npl lexels used by each notary. The font size of a lexel represents the frequency of tokens of that lexel in the notary's texts, and the colour of a lexel

represents the percentage of tokens of that lexel which occur with a zero inflection. The darker the lexel, the higher the percentage of zero tokens.

FIGURE 6.12: Word clouds showing the npl lexels used by each of the notaries public listed in table 6.6.

(A) Sir Thomas Crawfurd (B) Peter Marche (C) James Young

The wordclouds for Marche (figure 6.12b) and Young (figure 6.12c) appear fairly similar in the sizing and colouring of the various lexels, suggesting that the texts authored by these two notaries are likely to be similar in terms of content and subject. However, the wordcloud representing Crawfurd's texts (figure 6.12a) is different. The most frequent npl lexels in the texts of Marche and Young are *heir* and *land*, both with a low percentage of zero-inflection. Whilst Crawfurd's texts do display a high frequency of *heir*, the lexel which stands out clearly is *pertinent*, which is the most frequent npl lexel in Crawfurd's texts, and also has a high percentage of tokens inflected with zero. Nor is this high percentage of zero specific to Crawfurd - Marche's texts contain many fewer attestations of *pertinent* than Crawfurd's, but where this lexel does occur, it is likely to have a zero inflection.

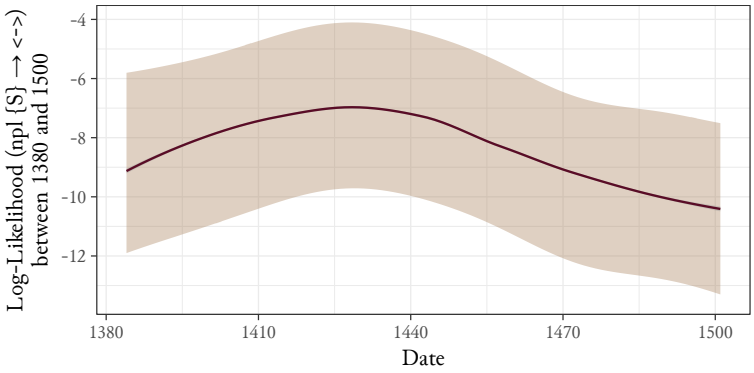
The apparent correlation shown in table 6.6, then, is an artefact of the high occurrence of a particular lexel which often occurs with a zero npl inflection. This same effect can also be seen with the lexel *witness*, which is attested more often in Crawfurd's texts, and is also particularly likely to be inflected with zero, as described in section 6.3.1. *Witness* is attested in the texts of both Marche and Young, where it occurs with a zero-inflection, but the high frequency of it in Crawfurd's texts gives the impression of his using zero-inflection more often.

6.3.2.4 Text date

Figure 6.13 shows the predicted likelihood of a zero realisation of npl {S} over the time period covered by LAOS. There is a weak trend shown whereby the likelihood of zero decreases in the second half of the fifteenth century, but the wide confidence bands shown around the trend line suggest that it is unwise to draw meaningful conclusions about any real fluctuation in the likelihood of zero over time.

This may seem contrary to the overall result shown by the summary of model 1, which indicated that

FIGURE 6.13: Estimated log-likelihood of zero for tokens according to text date estimated by Model 1.



date is a significant predictor of zero (χ^2 = 380.3; DF = 9, p < 0.01). However, this result indicates only that the inclusion of the date smooth in the model explains a significant amount of the observed variation in the dataset. Using a backwards-elimination technique to fit the optimum model, the decision to include date as a PV was taken on the basis that its inclusion improved the explanatory power of the model. A model identical to Model 1 but without date as a PV was found to be a worse fit to the data, meaning that less of the variation within the dataset can be explained without recourse to date as a PV. The confidence intervals in figure 6.13 indicate fairly low confidence in the specific trend shown by the smooth line, but the model results show that there is an effect of date which needs to be retained in the model to achieve the best fit.

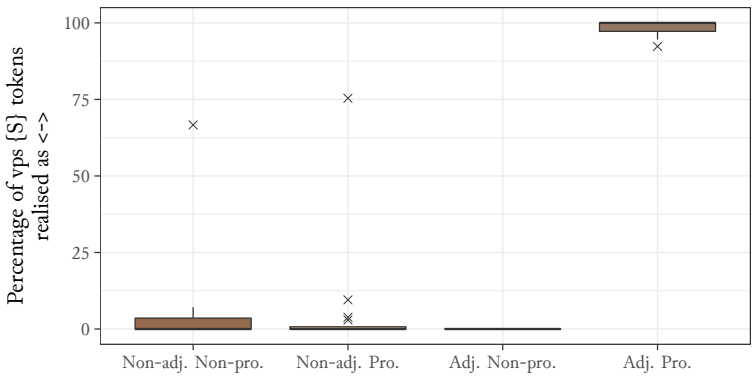
6.3.3 Zero-inflected vps tokens

As discussed in section 2.4.2.1, the realisation of vps {S} in OSc is conditioned by the NSR. The NSR captures a generalisation about OSc and Northern Middle English (NME): that the inflectional ending of first person singular and all plural vps forms is realised as zero or as <e> where the subject is (a) adjacent to the verb; and (b) a personal pronoun (de Haas 2011: 175).

Figure 6.14 shows the distribution of zero vps inflections in first person singular and all plural forms according to these two criteria. It is clear that the majority of zero-inflected vps forms occur where the subject is an adjacent pronoun, suggesting that the high proportion of zero inflection displayed by vps forms as shown in table is accounted for by the operation of the NSR. The three environments which do not fulfil both requirements for the NSR have extremely low levels of zero, with the exception of tokens of *oblige*. The reason for a high proportion of zero for *oblige* is not intuitively clear, but an examination of the orthographic forms of *oblige* in INFLAOS reveal that they are most often realised with a final <s> or <ß>. *Oblis* is described by the DSL as a variant of *oblige* more common in OSc.

Given the clear correlation between vps zero and the combination of subject pronominality and adjacency, the GAM described in section 6.3.4 includes a single PV, 'NSR environment', a binomial variable

FIGURE 6.14: Realisation of first person singular and all plural vps tokens as zero by adjacency and pronominality of subject.



with the following categories:

(a) NSR environment - this category is assigned to any vps token, the subject of which is:

1. a personal pronoun, identified by the tag '<P' appended to the grammel; and
2. adjacent to the vps token, identified by a '+' appended to the grammel.

The full grammel for tokens categorised as being in an NSR environment is vpsxy<P+, where x is either 1 (denoting a singular subject) or 2 (denoting a plural subject); and y is either 1, 2 or 3, denoting the grammatical person of the subject. For example, (19) has the grammel vps11<P+, indicating a present-tense verb with an adjacent first-person singular personal pronoun subject. This token would be classified as 'NSR environment'

(b) non-NSR environment - this category is assigned to any token which does not fit the criteria in (a), such as (20), which has the grammel vps23<n+, indicating a present-tense verb with an adjacent third-person plural noun subject; and (21), with the grammel vps21<P-, indicating a present-tense verb with a non-adjacent first-person plural personal pronoun subject.

(19) <disasent> vps11<P+ *disassent* 'to refuse assent' (*Disassent, v.* 2004) [text 1409: 1459, CAR, AGS]

(20) <schawis> vps23<n+ *shows* [text 723: 1495, NOT, MLO]

(21) <adnulſ> vps21<P- *annuls* [text 360: 1441, CHA, WLO]

Figure 6.15 shows the percentage of vps tokens from each lexel in INFLAOS with a zero inflection according to stem-final *littera*. The mean values of <e> and <ß> stand out as considerably higher than the means for other SFL. As shown by figure 6.14, the realisation of vps {S} as zero appears strongly correlated with NSR environment. The fact that figure 6.15 shows a high percentage of zero-inflection in tokens with

stem-final <e> fits with the suggestion made by de Haas (2011) that the operation of the NSR causes both zero and <e>-final forms of vps {S}.

FIGURE 6.15: The percentage of vps tokens from each lexel in INFLAOS with a zero inflection according to stem-final *littera*.

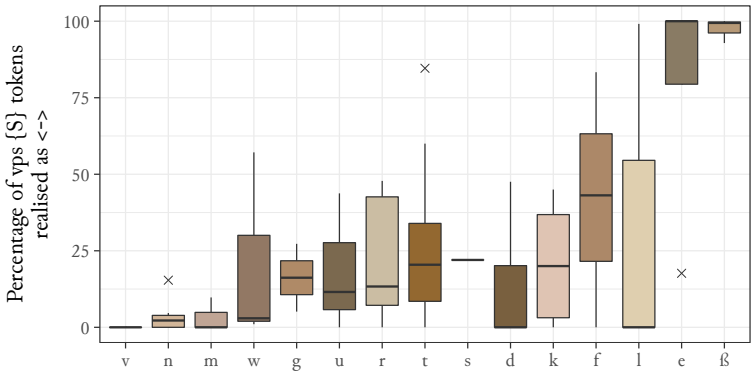


Figure 6.16 shows the percentage of vps tokens of each lexel in INFLAOS with a zero inflection over time. As in figure 6.6, the straight red line and shading represents a linear trend line fit to the individual lexel data points and the corresponding 95% confidence interval, and the blue line and shading represents a smooth trend line and the corresponding 95% confidence interval fit to the same points. The data points are much sparser for vps than for npl due to npl tokens being much more common in the corpus. Whereas in the npl plot it seemed that there might be some slight positive correlation between date and realisation of {S} as zero, there is no evidence shown in this plot for a correlation between date and vps {S} being realised as zero. This may be due to the fact that alternation between fully-realised forms of {S} and zero is a regular part of the OSc vps paradigm.

FIGURE 6.16: The percentage of vps {S} tokens of each lexel in INFLAOS realised as zero between 1380 and 1500. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Red: linear trend line; blue: smooth trend line.

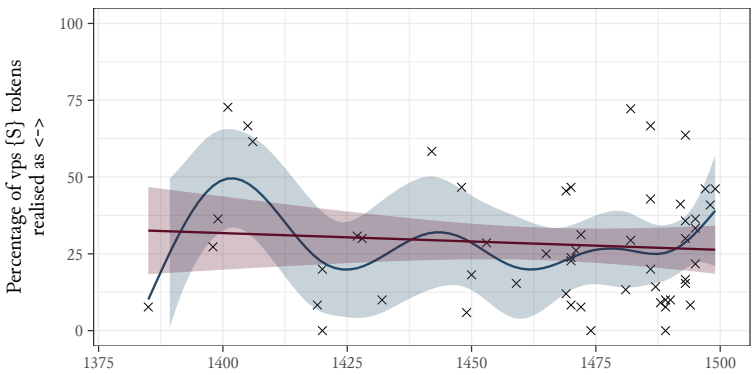
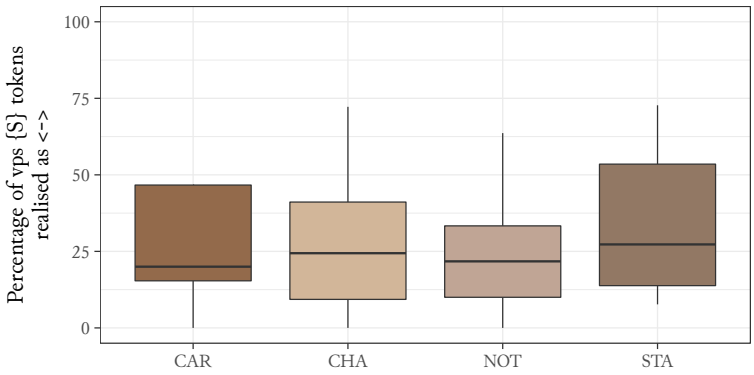


Figure 6.17 shows the percentage of vps tokens from each text in INFLAOS with a zero inflection

according to text type. The most noticeable feature of this series of box plots in comparison with figure 6.7 which shows the percentage of zero npl tokens per text is that there are no outliers in this plot. In contrast, there were outlying cartularies, state documents, charters and burgh records for npl. In addition, the mean percentage values for vps are higher than for npl, indicating a higher prevalence of zero vps tokens than zero npl tokens for all text types. The vps box plots for all text types also display much wider overall and interquartile ranges than their npl counterparts, which suggests that vps zero is more widely dispersed throughout the corpus of texts, rather than confined to a few. Having said this, it should be noted that figure 6.17 does not contain burgh record plot for vps because no burgh record texts have more than 10 vps tokens. This reflects the overall lower prevalence of vps tokens than npl tokens in INFLAOS, combined with the fact that, as shown in section 4.1.5.2, burgh records are on average much shorter than texts of other types.

FIGURE 6.17: The percentage of vps tokens from each text in INFLAOS with a zero inflection according to text type.



6.3.4 Modelling the likelihood of zero realisation of vps {S}

Compared to the equivalent model for npl {S}, the GAM of vps zero inflection shows that a smaller number of PVs have a significant effect on the likelihood of vps {S} being realised as zero. Whereas some contextual information (individual text, date and type) showed a significant effect on the realisation of npl {S} as zero, these factors do not appear significant in the vps model. Instead, the list of significant PVs is reduced to the random effect of individual lexel; whether or not a token is susceptible to the operation of the NSR; and the stem-final *littera*.

Table 6.7 presents the result of an ANOVA comparison between Model 2 and another model which includes the same PVs as Model 2, but with the addition of a PV encoding the grammatical person and number classification of each token. The ANOVA table compares the two models with the purpose of ascertaining whether the more specific model accounts for the variation in the dataset significantly better than the more general model. In this case, the model including the extra PV of person and number is the

Model 2: The results of a generalised additive model of zero inflection in INFLAOS vps tokens.
R^2 = 0.945; Deviance explained = 92.4%; N = 1,647

Parametric terms

		Estimate	SE	z	
(Intercept)		-6.25	1.17	-5.35	***
SFL	e	6.26	1.6	3.9	***
	f	3.48	1.94	1.79	.
	g	3.53	1.6	2.21	*
	k	-0.83	1.35	-0.62	
	l	3.47	1.56	2.22	*
	m	0.36	2.08	0.17	
	n	-2.18	1.47	-1.49	
	r	2.59	1.47	1.75	.
	s	-1.72	1.57	-1.09	
	ß	10.92	1.85	5.9	***
	t	0.08	1.24	0.07	
	u	-0.58	1.83	-0.32	
	w	3.72	1.47	2.54	*
NSR env.	Yes	9.26	0.83	11.16	***

Approximate significance of smooth terms

	edf	Ref.df	χ^2	
Lexel	34.66	202	85.84	***

more specific model. Generally, we expect the more specific model to be a better fit to the data, simply because it includes more explanatory variables. The ANOVA comparison takes into account the increased complexity of the model together with the level of improvement in accounting for the data this increase in complexity provides (the increase in 'explained deviance'). The summary of Model 2 shows that the model including person and number as well as NSR explains more deviance than the model containing NSR only (an increase in explained deviance of 9.09 between the two models) . This increase in explained deviance comes with an increase in complexity of 4.08 degrees of freedom. Overall, the ANOVA test determines that the model including person and number does not do a significantly better job of explaining the variation in the data than the model containing only NSR, as shown by the p-value of >0.01.

Table 6.7: ANOVA and Akaike Information Criterion (AIC) comparison of Model 2 and another model which includes the same PVs as Model 2, but with the addition of a PV encoding person and number. The result of the ANOVA comparison indicates that the difference between the two models is not significant: p>0.01.

	ANOVA comparison		AIC comparison	
Model	**Residual DF**	**Residual Deviance**	**DF**	**AIC**
Model 2	1,577.40	170.98	50.75	272.48
Model 2 + pers./num.	1,573.30	161.90	57.07	284.51
Difference	**4.10**	**9.08**	**-6.32**	**-12.04**

The fact that adding the effect of person and number into the model does not provide a statistically significant improvement in the model's ability to account for the data suggests that the application of the

NSR is prevalent in the INFLAOS data, as there is clearly an alternation between the non-NSR paradigm with its ubiquitous {S} and the zero forms which characterise the NSR paradigm. To further confirm this, table 6.8 shows another ANOVA comparison, this time between the model containing PVs NSR and person and number, and a model containing person and number but no NSR PV. In this case, the ANOVA shows that omitting NSR as a PV from the model causes a considerable decrease in the deviance explained (the model including NSR as a PV has an explained deviance figure 839.39 higher than the model which does not include NSR). This ANOVA comparison does not return a p-value. This is because the purpose of the p-value is to estimate the statistical significance or lack thereof of the difference between two models. In effect, it indicates which model is better by trading off complexity against explanatory power. In this case, the model without NSR is actually judged to be both more complex and less explanatory. No p-value is necessary because there is no basis on which the model without NSR could be judged better.

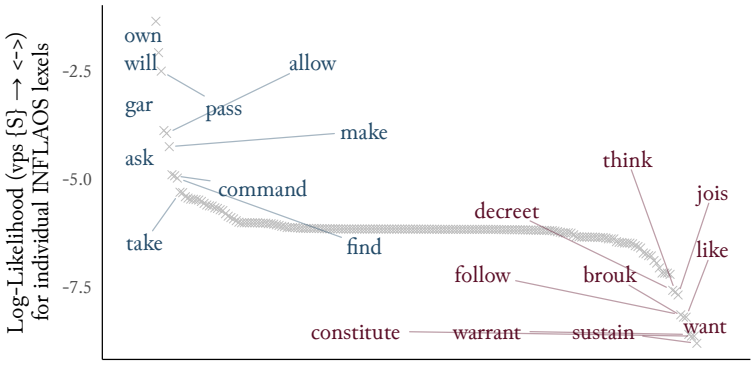
Table 6.8: ANOVA and AIC comparison of a model containing PVs NSR and person and number, and a model containing person and number but no NSR PV.

	ANOVA comparison		AIC comparison	
Model	**Residual DF**	**Residual Deviance**	**DF**	**AIC**
Model 2 + pers./num.	1,573.30	161.90	55.09	272.08
Model 2 - NSR	1,519.10	1,001.30	95.30	1,191.89
Difference	**54.20**	**-839.40**	**-40.21**	**-919.81**

6.3.4.1 Individual lexels

Figure 6.18 shows the estimated log-likelihood of individual lexels to have a zero vps inflection. Lexels which appear higher on the y-axis are estimated by the GAM to have an unusually high log-likelihood of zero considering all other significant factors in the model. The plot therefore visualises the random variation in the GAM caused by individual lexels. The top and bottom 10 lexels are labelled.

Figure 6.18: Fitted values for the random effect of individual lexel in a generalised additive model showing the significant effects of etymology, SFL and NSR conditions on the realisation of vps morphemes as zero in INFLAOS.



The most significant outlier in figure 6.18 is *own*. Qualitative analysis of the data show that vps *own* is attested in INFLAOS seven times, five 3pl. forms and two 3sg. forms, all in non-NSR environments. The lack of a realised {S} inflection in these tokens is explained by the fact that they are not forms of the PDE word *own*, but rather of *owe*, realised as <aw> (5 tokens), <acht> (1 token) and <aht> (1 token). The DSL lists this sense of *owe*, meaning 'to own' as a dialect form still in use in the early twentieth century (*Awe v.1* 2004), but the bulk of attestations are from OSc texts dated before 1650 (*Aw v.* 2004). These demonstrate the general use of the uninflected form *aw* for all grammatical categories of vps, though the entry acknowledges the tendency for past tense form to "supplant the present" as with the English form *ought*, originally the past tense form of *owe*.

The next most extreme outlier after *owe* is *will*. There are 116 tokens of vps *will* in INFLAOS, all but one of which have a zero inflection. Of these zero-inflected tokens, 92 (80%) are in a non-NSR environment, or are unaffected by the NSR (3sg., 83 tokens). It is clear from these figures why *will* is judged to be an outlier. Unlike the qualitative analysis of *own*, there is nothing which immediately suggests why this lexel should be attested almost solely with zero inflection. However, a concordance of all tokens of vps *will* in their textual contexts reveals that 54 tokens form part of the phrase *as [the] law will*. Other variants on this phrase, with law as the subject of the clause, are also attested. Table 6.9 lists a total of 78 such phrases. The DSL includes this sense of *will* as a variant on its main usage as an "expression of an authoritative intention" (*Wil(l, v.1* 2004). Specifically, the use of *will* where the object of the clause is omitted, which occurs most often in the *as law will* construction shown in table 6.9. The use of zero-inflected vps *will* as part of this specific phrasing appears likely to account for its appearance as an outlier in figure 6.18.

Table 6.9: INFLAOS tokens of vps 3sg. *will* as the verb in a clause where the subject is *law*.

Phrase including *law will*	**Tokens**
as [the] law will	54
as [the] course of common law will	7
at law will	4
otherwise than the course of common law will	4
as law of [the] march[es] will	2
[?] law will	1
& law will	1
as common law will	1
as the order of law of Scotland will	1
otherwise than law will	1
that law will	1
when and where law will	1
Total	**78**

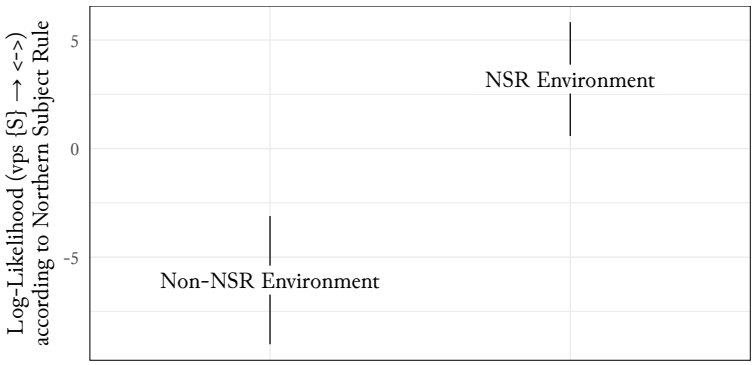
Another outlier in figure 6.18 is *gar*. There are eleven instances of *gar* in INFLAOS, of which six tokens (in six different texts) have a zero inflection. At first glance, it seems that this may be a localised usage, as

four of the texts are cartularies from the county of Angus. The other two are acts of parliament localised to Edinburgh though, interestingly, one of them concerns the "swearing in of Archibald, earl of Angus, as warden of the East March" (Williamson 2008). Having said that, closer inspection reveals that three of the cartulary texts are in fact copies of a single original document, a letter by Robert, Duke of Albany. In the first two copies, *gar* is realised as <gar>, and in the third copy as <gare>. These are the only tokens of *gar* which have a medial <a> - both the zero and non-zero-inflected forms have medial <e>.

6.3.4.2 NSR

Figure 6.19 shows the log-likelihood of vps {S} tokens being realised as zero in NSR and non-NSR environments (see section 2.4.2.1 for an explanation of the NSR, and the discussion of figure 6.14 for the representation of NSR environments in LAOS). It is clear that it is more likely for npl {S} to be realised as zero in NSR than in non-NSR environments.

FIGURE 6.19: Estimated log-likelihood of zero for vps {S} tokens in NSR and non-NSR environments.

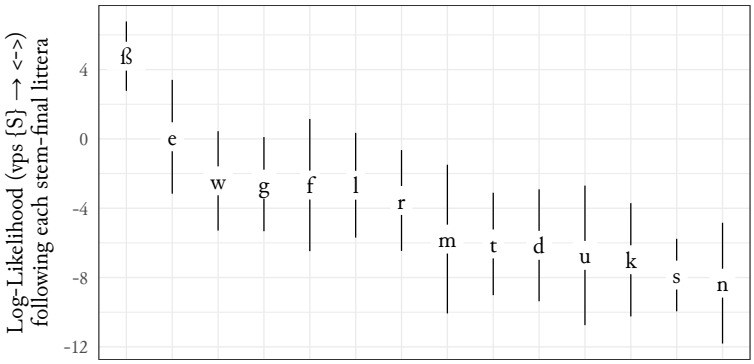


6.3.4.3 stem-final *littera* (SFL)

Figure 6.20 shows the estimated log-likelihood of vps {S} being realised as zero following various SFL. As suggested by the descriptive boxplot of percentage zero by SFL presented in figure 6.15, stem-final <ß> stands out from the rest as highly likely to be followed by a zero realisation of vps {S}. However, whilst the descriptive plot showed that stem-final <e> occurred almost as frequently preceding zero as stem-final <ß>, the log-likelihood plot implies less of a similarity between these two SFL in this regard. This suggests that, despite the similar percentage of zero vps inflection exhibited by lexels with stem-final <e> and <ß>, when all other factors in the model are taken into account, the apparent effect of stem-final <e> as a predictor of zero is in fact explained by other factors.

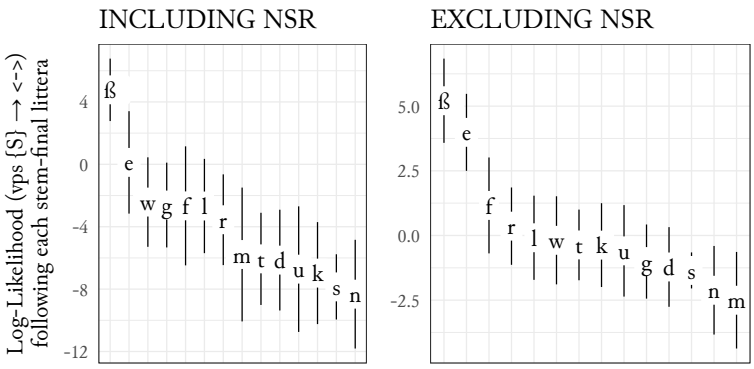
Figure 6.21 shows two plots of the correlation between SFL and log-likelihood. The left-hand plot shows the coefficients for each SFL as estimated by Model 2 (including NSR as a PV), and the right-hand

FIGURE 6.20: Estimated log-likelihood of zero for tokens according to SFL.



plot shows the coefficients for each SFL estimated by another GAM which is identical to Model 2 except that the NSR PV is not included. The comparison of these two plots shows that including NSR as a PV renders the correlation between stem-final <e> and the log-likelihood of zero insignificant. This supports the idea that stem-final <e> acted as an inflectional marker in NSR environments (de Haas 2011: 175) .

FIGURE 6.21: Comparison of log-likelihood of zero for tokens according to SFL in Model 1 including the effect of NSR environment and a GAM excluding the effect of NSR environment.



Chapter 7

Scribal Abbreviation

7.1 Introduction

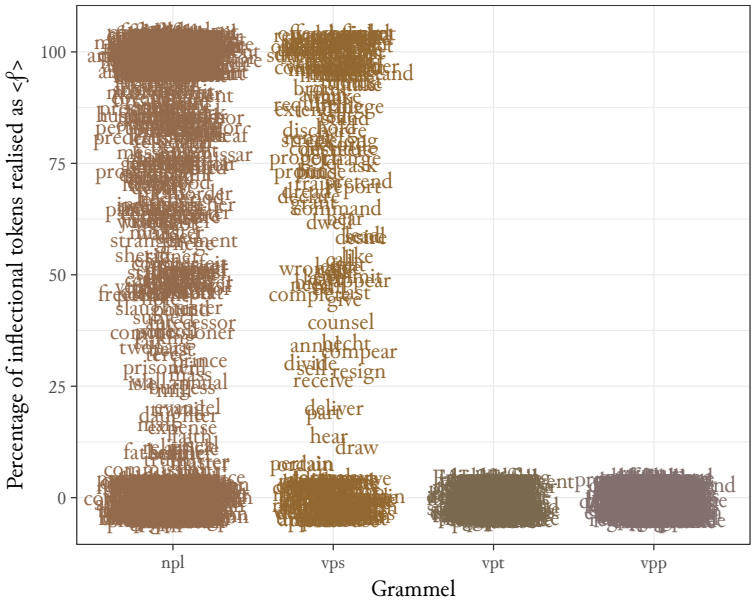
This chapter presents the correlations between the predictor variables (PVs) introduced in section 4.1.5 and the dependent variable (DV) of abbreviation, and investigates (a) whether the orthographic representation of {S} as <ſ> is motivated solely by palaeographic convenience; or (b) whether its use correlates with other factors such as spatio-temporal variation, text type, or other lexical variables.

Figure 7.1 shows, for each grammel, all lexels in *Inflections in A Linguistic Atlas of Older Scots* (INFLAOS) exemplified by 10 or more non-zero tokens. The position of each lexel on the y-axis indicates the percentage of tokens of that lexel where the inflection is realised as <ſ>. All past tense verb (vpt) and past participle (vpp) lexels are clustered around 0% as <ſ> is only attested with plural noun (npl) and present tense verb (vps). There are two main clusters of lexels for both npl and vps at the extremes of the percentage scale, with a smaller amount of lexels between these two extremes. This distribution suggests that there may be a binary distinction to be found among the predictors of <ſ>.

7.2 Abbreviation of npl and vps {S}

Figure 7.2 shows boxplots of individual text and lexel percentages of npl and vps {S} realised as <ſ>. Each pane contains a boxplot for each of npl and vps. The left pane plots percentage <ſ> by individual lexel, and the right by individual text. Whilst the overall ranges of all of these plots have similar upper and lower extents, the interquartile ranges of the percentage values of individual lexels are much larger than for individual texts. This suggests that the variation between lexels is greater than that between texts. In other words, individual texts are more likely to be similar to one another in terms of percentage of abbreviation than individual lexels. The lexel boxplots for npl and vps are very similar, but the text boxplots for each

FIGURE 7.1: Realisation of npl, vps, vpt and vpp inflections as <ſ>. Only lexels with >10 tokens included.



grammel differ, with a lower mean percentage vps <ſ> per text. This suggests that intertextual variation is more of a factor for vps <ſ> than for npl <ſ>. Specifically, certain scribes may have a tendency to use <ſ> less than others, whereas there is no such equivalent generalisation indicated for individual lexels.

FIGURE 7.2: The percentage of npl and vps {S} tokens realised as <ſ>. The left-hand plot is based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plot is based on the percentage by text (so individual data points represent individual texts).Only texts and lexels with more than 10 tokens are included.

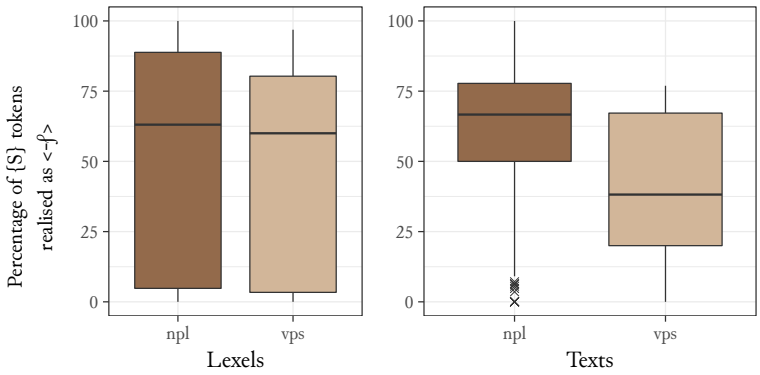
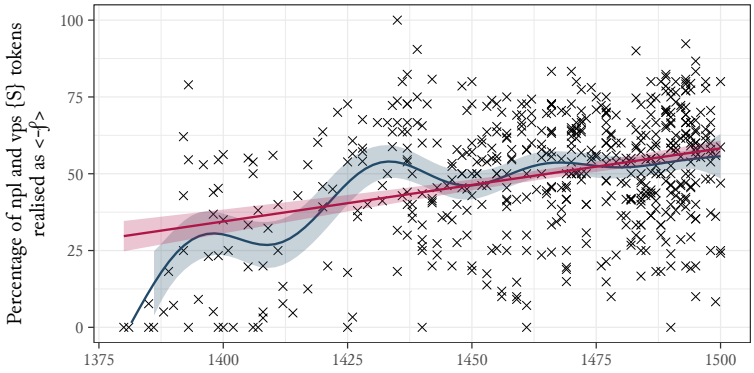


Figure 7.3 shows the percentage of npl {S} tokens realised as <ſ> over time. Each point represents an individual text, and only texts containing more than 10 npl tokens in INFLAOS are shown. The blue line represents the percentage of abbreviated tokens as a smooth function of the PV date, and the red line represents the same relationship modelled using a linear function. The shaded areas around each line represent 95% confidence intervals.

FIGURE 7.3: The percentage of npl and vps {S} tokens of each lexel in INFLAOS realised as <ſ> between 1380 and 1500. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Red: linear trend line; blue: smooth trend line.



The confidence intervals of both the linear and smooth functions are widest for the period 1380-1400, getting narrower as the value of date increases. This indicates a lesser degree of certainty in the trend for earlier dates due to lower text numbers (see section 4.1.5). Whether plotted as a smooth or a linear function, the percentage of npl and vps {S} tokens realised as <ſ> clearly increases over time. The smooth line, however, reveals a large amount of fluctuation in abbreviation percentage per text in the first half of the period, in contrast to the second half of the period, where the smooth line follows the linear line much more closely. The fluctuation in the percentage of abbreviation in the first half of the period is in part due to the lower frequency of texts, which leads to the smooth line being more influenced by outlying texts than in the latter half of the period. However, the distribution of individual points suggests that this trend is legitimate. Most noticeably, texts with no abbreviation of {S} at all are all dated before 1425. After this date, all texts have at least one <ſ>, with the majority clustering above the 50% mark.

Figure 7.4 shows the percentage of npl and vps {S} tokens realised as <ſ> in each text, arranged into box plots by text type. The series of boxplots shows the variation in abbreviation levels in texts of each type. The category which stands out is state documents (STA). The upper and lower quartiles and the median of this boxplot are considerably lower than those of the other four categories, indicating a lower overall percentage of abbreviated tokens. The upper whisker of this boxplot does not extend as high as those of the other categories, with the exception of cartularies (CAR). This indicates that no cartularies or state documents use {S} abbreviation exclusively, in contrast to charters (CHA) and burgh records (BUR). Having said this, the cartulary boxplot is shorter than the state documents boxplot, with the former's lower whisker not extending to 0% as the latter's does. The interquartile range of the cartulary box plot is also smaller, suggesting less variability in the presence or absence of abbreviations between different texts. The boxplot for charters extends to 100% at the extreme of the upper whisker, and to 0% including outlying data points (labelled with text numbers).

FIGURE 7.4: The percentage of npl and vps {S} tokens realised as <ſ> in burgh records (BUR), cartularies (CAR), charters (CHA), notarial protocol books (NOT) and state documents (STA). Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown.

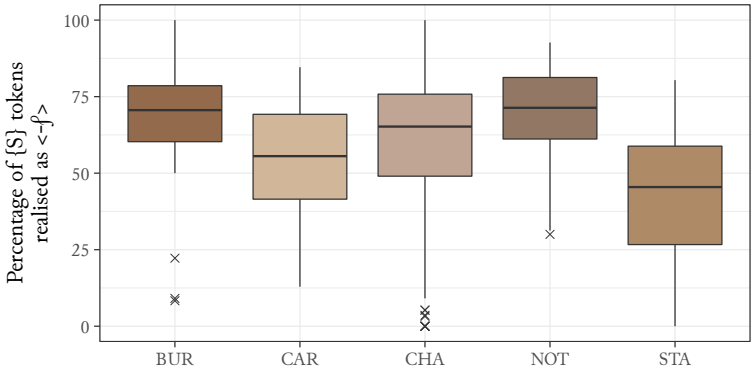
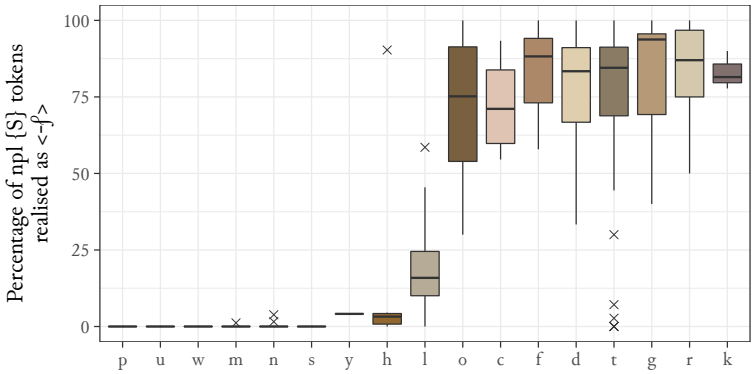


FIGURE 7.5: Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown.



This phenomenon is recorded in *A Linguistic Atlas of Older Scots* (LAOS) as an abbreviation of {S} where it occurs with a plural noun. Without examining each manuscript, it is impossible to say definitively how often, if at all, stem-final <l> is followed by <ſ> rather than a cross-stroke through a double-<l>. However, table 7.1 shows the percentage of npl {S} abbreviation following single and double stem-final <l>. Single stem-final <l> is followed by an abbreviation in 11% of cases, whilst double stem-final <l> is followed by abbreviation in 41% of cases. The much higher prevalence of abbreviation recorded for stem-final double <l> suggests that the cross-stroke, which is often attested with stem-final <ll> in both singular and plural nouns, may account for a significant amount of the npl {S} abbreviation following stem-final <l> in INFLAOS.

FIGURE 7.6: A cross-stroke through stem-final <ll> in <a(n)nuall(is)> *annuals* [text 1645, 1445, BUR, ABD].

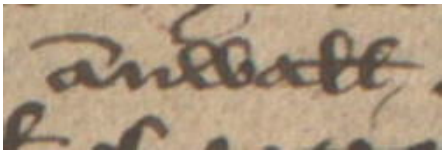
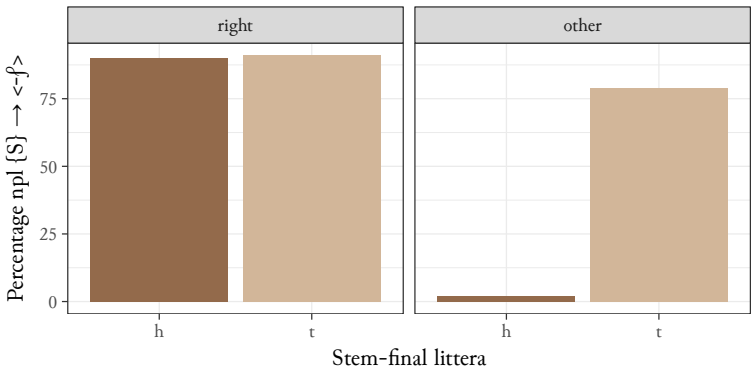


TABLE 7.1: The percentage of npl {S} abbreviation following single and double stem-final <l>

Stem-final <l>	Not Abbreviated	Abbreviated	Total
Single	293 (89%)	36 (11%)	329 (100%)
Double	102 (59%)	72 (41%)	174 (100%)
Total	395 (79%)	108 (21%)	503 (100%)

Stem-final <m> and <n> have outlying points, but since there is so little variation in the levels of abbreviation for these stem-final *littera* (SFL), the lexels *term*, *condition* and *person* are classified as outliers despite a very low level of abbreviation. On the other hand, <h>-final stems are not generally followed by abbreviation with the exception of the lexel *right*. This is interesting, because *right* is written with a final <t> in the majority of instances in INFLAOS, as well as being phonetically [t]-final. Figure 7.7 shows the percentage of npl {S} tokens following stem-final <h> and <t> where those *littera* occur at the end of *right* and where they occur at the end of other lexels. With *right*, abbreviation occurs in almost all cases whether the SFL is <h> or <t>, despite the extremely low prevalence of abbreviation following stem-final <h> in other lexels.

FIGURE 7.7: The percentage of npl {S} tokens with stem-final <h> and <t> realised as <ſ>: *right* compared with all other lexels.



Stems with final <t>, as shown by figure 7.5, have a high overall level of abbreviation. SFL <t> has several outlying points, in this case, lexels with much lower levels of abbreviation than the majority of <t>-final stems. With the exception of *indenture*, all of these outlying lexels end in *-er*: *charter*, *daughter*, *master*, *matter* and *water*. This suggests that the corresponding manuscript forms of the tokens of these lexels end in a siglum signifying <er>, <ꝛ> following stem-final fully-realised <t>.

Aside from the obvious appearance of two distinct groups of SFL in INFLAOS with regard to abbreviation, it is notable that the SFL at the lower end of the percentage scale exhibit considerably less variation in range and interquartile range of percentage scores than those at the higher end of the scale. Those at the lower end of the scale are represented by horizontal lines, indicating that the percentage of abbrevia-

tion for lexels ending in these SFL is extremely homogenous. By contrast, those at the higher end of the scale exhibit a much wider range of percentage abbreviation between different lexels. For example, lexels with stem-final <n> have a mean percentage abbreviation extremely close to 0%, as well as a range and interquartile range extremely close to 0%. Conversely, lexels with stem-final <d> have a mean percentage abbreviation of around 80%, but with percentage values for individual lexels ranging from around 30% to almost 100%. This suggests that there are other factors influencing the use of abbreviation, but that abbreviation is confined to a certain set of SFL.

7.2.1 Modelling the likelihood of <ſ> abbreviation in npl and vps {S}

Model 3 shows the results of a generalised additive model of the effects of SFL, date, location and text type on likelihood of npl and vps {S} morphemes being realised as <ſ> in INFLAOS. The final model formula which includes only PVs found to be significant using a backwards-elimination technique (see section 5.2.2). Significant contextual PVs are text date and type. Text date is included in the model as a smooth term, enabling the model to capture a non-linear trend in the likelihood of abbreviation over time; and text type is a categorical variable. The significant lexical PVs are SFL and grammel. SFL categories with significant p-values (indicated by asterisks in the summary table) are significantly different from the reference category <d> in terms of their likelihood of being followed by <ſ>. All the significant SFL shown in summary table 3 have negative coefficient values, meaning that the likelihood of their being followed by <ſ> is significantly lower than the likelihood of stem-final <d> being followed by <ſ>. Individual text and lexel are included as random effects in the model. As explained in section 5.2.5, this means that they are not of direct interest to the investigation as PVs, but they are an important source of potential variance in the data which needs to be accounted for if we are to have confidence in the model's estimates of the effects of the fixed PVs.

In the summary of Model 3, the statistic indicating overall fit of the model is explained deviance, which is 66.6%. This is the approximate amount of the variation in the dataset which the model can successfully account for. A figure of 66.6% indicates that the model is able to accurately predict a considerable amount of the variation in the data using the combined explanatory power of the PVs listed above, but there is still a fairly large amount of variation in the data which is not explained by these predictors. This indicates that there are factors affecting the likelihood of abbreviation which are not captured by the model.

In terms of individual PVs, Model 3 contains only those predictors found to be statistically significant. All the predictors remaining in the model are significant at $p<0.01$. There are two significant random effects, individual text and individual lexel. This means that there is a significant amount of variation in the npl INFLAOS data which can be attributed to idiosyncratic differences between individual words or individual texts. However, the purpose of including these random effect terms in the model is not to investigate them

MODEL 3: **R^2 = 0.706; Deviance explained = 66.6%; N = 12,731**

Parametric coefficients:

		Estimate	SE	*z*	
(Intercept)		1.34	0.92	1.46	
SFL	c	0.44	0.53	0.83	
	f	-0.02	0.48	-0.04	
	g	0.27	0.35	0.78	
	h	-4.14	0.46	-8.93	***
	k	1.17	0.40	2.94	**
	l	-3.00	0.31	-9.82	***
	m	-8.32	0.87	-9.53	***
	n	-6.40	0.40	-16.09	***
	o	-2.15	0.31	-6.90	***
	r	0.39	0.24	1.62	
	s	-5.35	0.65	-8.24	***
	t	-0.22	0.23	-0.95	
	u	-6.95	1.11	-6.27	***
	w	-7.51	1.10	-6.83	***
	y	-3.97	1.25	-3.18	**
Type	BUR	0.67	0.17	3.99	***
	CAR	-0.54	0.19	-2.77	**
	NOT	0.31	0.23	1.35	
	STA	-0.95	0.25	-3.88	***
Grammel	vps	-0.39	0.18	-2.23	*

Approximate significance of smooth terms:

	EDF	Ref. DF	χ^2	
Text	430.66	1,052.00	2,331.28	***
Lexel	244.10	925.00	1,316.59	***
Date	5.46	5.89	75.98	***

for their own sake - there is little benefit to explaining variation at the level of individual texts or words if there is no overarching variation - but to ascertain whether or not the fixed effect PVs remain significant when individual textual and lexical variation is accounted for. This involves firstly ascertaining whether the individual textual and lexical differences are significant. This is done by comparing models fit with and without the random effect term. The significant p-values for both text and lexel indicate that the effects of these PVs explain a significant amount of the random variation in the dataset. The significant fixed effects in the model are estimated to be significant even when these individual differences are taken into account. For example, the inclusion of the significant random effect of lexel in the model alongside the significant fixed effect of SFL shows that the variation accounted for by SFL is not an artefact of the variation between different lexels, such as a high likelihood of abbreviation on certain lexels beginning with certain SFL.

Table 7.2 lists comparative Akaike Information Criterion (AIC) scores for Model 3, as well as nested models, each with one PV removed. AIC is a statistic which measures the quality of regression models fit to the same dataset relative to one another. The 'quality' of each model is assessed according to the variation

in the dataset which it accounts for, traded-off against its complexity. The larger the difference between the AIC score of Model 3, which includes all PVs found to be significant, and the AIC score of a nested generalised additive model (GAM) with one PV removed, the larger the improvement in the accuracy of the GAM caused by the predictor which has been removed. Removing the random effect of text from Model 3 causes the largest AIC difference, suggesting that this factor makes the biggest improvement to the model in terms of how much of the variation in the data it accounts for.

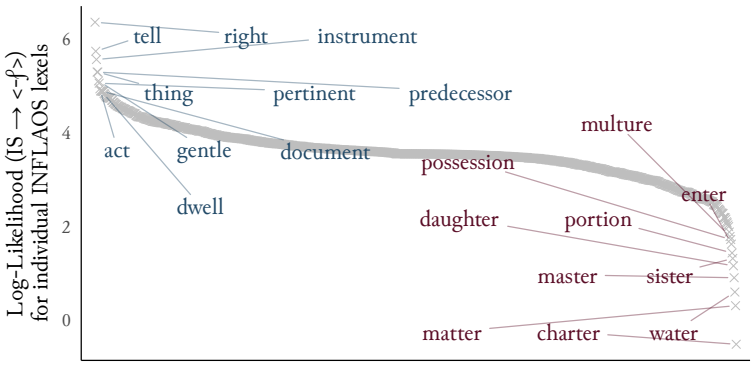
TABLE 7.2: A comparison of AIC scores for Model 3 and 6 other models fit by consecutively removing single PVs from this model.

Full model	PV removed	DF	AIC	Δ AIC
Model 3		705	7,171	
	Text	294	8,525	1,354
	SFL	1,003	8,215	1,045
	Lexel	469	7,875	704
	Type	726	7,228	57
	Date	728	7,211	40
	Grammel	705	7,174	3

7.2.1.1 Individual lexels

Figure 7.8 shows the log-likelihood of npl and vps tokens of a particular lexel having <ſ> as estimated by Model 3. That is, taking into account all the other significant predictors in the model, some lexels appear unusually likely to have <ſ>, whereas others are unusually unlikely to have <ſ>.

FIGURE 7.8: Fitted values for the random effect of individual lexel in a generalised additive model showing the significant effects of SFL, grammel, date and text type on likelihood of npl {S} morphemes being realised as <ſ> in INFLAOS.



As suggested by the percentage plot of SFL in figure 7.5, the lexel *right* is the uppermost outlier. This reflects the tendency of scribes to use an abbreviation with this lexel following both stem-final <h> and stem-final <t>, despite the lack of abbreviation following stem-final <h> elsewhere. Other outliers at the

higher end of the log-likelihood scale include *tell*, *dwell* and *gentle*. Both of these are generally spelled with stem-final <ll> which, as noted in the discussion of figure 7.5, often had a cross-stroke which is recorded as an abbreviation in LAOS.

Other outliers at this end of the scale include lexels typically realised with 'abbreviation-friendly' SFL, such as *act* and *knight*, which typically have stem-final <t>. Because the model predicts the occurrence of abbreviation, and therefore the lexels which are unusually likely or unlikely to display it, using a combination of all the PVs included in it, it is not possible to pinpoint precisely why a given lexel appears at a particular point on this plot. In the case of *right*, the initial descriptive analysis had already suggested that this lexel might exhibit different behaviour. Not so for *act*, which does not appear as an outlier for <t> in figure 7.5. However, it is possible to suggest a reason for this by considering the other significant PVs in the model together with the tokens of npl *act* in INFLAOS. Specifically, 14 out of a total 24 tokens of *act* occur in state documents, and 11 out of those 14 have an abbreviated inflection. Model 3 showed that state documents were significantly less likely than other text types to contain npl {S} abbreviation ($z = -4.59$; $p < 0.01$), so the fact that *act* is estimated by the model as unusually likely to occur with abbreviated {S} may well be due, at least in part, to the fact that the texts in which it often occurs generally contain less npl {S} abbreviation. The model does not account for the semantic associations of lexels, or for links between texts beyond the four types specified in the model. As a human interpreter, however, it seems fairly unsurprising that *act* is unusually likely to be abbreviated in state documents, particularly as an examination of the LAOS county codes for the six documents in which npl *act* occurs shows that they are all *acts* of parliament.

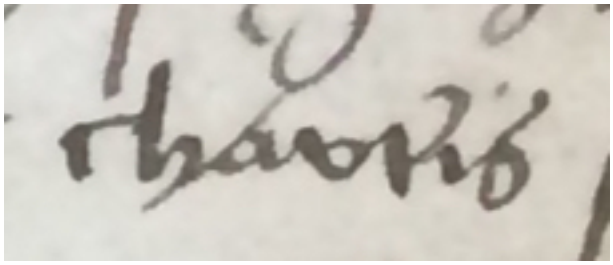
The lower end of the log-likelihood scale in figure 7.8 is dominated by the same {ER}-final lexels which were shown to be outliers for stem-final <t> in figure 7.5. Table 7.3 shows the percentage of abbreviated and non-abbreviated npl {S} tokens with stem-final <t>, according to the type of siglum appearing after the SFL (the stem-final *siglum* (SFS)).

TABLE 7.3: The percentage of abbreviated and non-abbreviated npl {S} tokens with stem-final <t>, according to SFS.

Stem-final siglum	**Not abbreviated**	**Abbreviated**	**Total**
(er)	138 (98%)	3 (2%)	141 (100%)
(ur)	11 (69%)	5 (31%)	16 (100%)
none	245 (14%)	1466 (86%)	1711 (100%)
horizontal flourish	0 (0%)	3 (100%)	3 (100%)
Total	**394 (21%)**	**1477 (79%)**	**1871 (100%)**

When no SFS intervenes between stem-final <t> and npl {S}, {S} is abbreviated 86% of the time. When stem-final <t> is followed by <ʾ>, however, {S} is abbreviated in only 2% of cases. Figure 7.9 shows an example of stem-final <tʾ> in the word *charters*. The <ʾ> symbol is formed by extending the pen-stroke upwards after forming the cross-stroke of the <t>. The cross-stroke is therefore not available to facilitate the abbreviation of {S} as <ſ>, and the inflection is consequently realised fully as <is>.

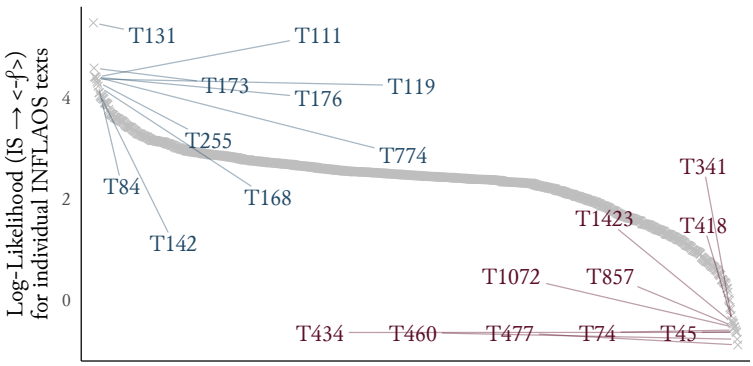
FIGURE 7.9: <chart(er)is> *charters* [text 7: 1493, CHA, AYR].



7.2.1.2 Individual texts

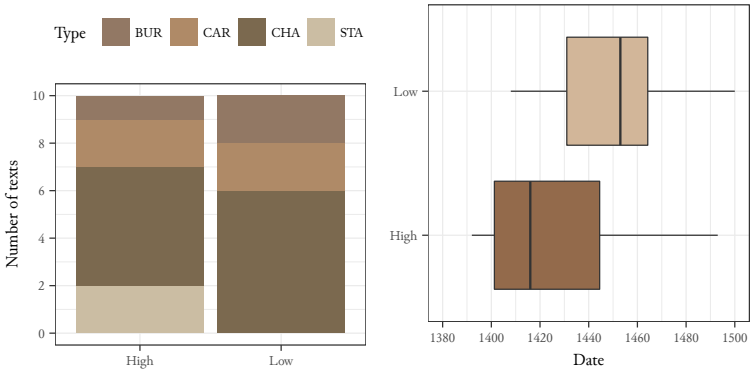
Figure 7.10 shows the fitted values for the random effect of individual text in Model 3. As with the previous plot of lexel as a random effect, the text numbers placed highest on the plot indicate those which are unusually likely to have abbreviated tokens considering their categorisation with regard to the significant PVs in the model.

FIGURE 7.10: Fitted values for the random effect of individual text in a generalised additive model showing the significant effects of SFL, date, location and text type on likelihood of npl and vps {S} morphemes being realised as <ſ> in INFLAOS.



It is not as intuitive a task to pick out and compare outlying text numbers as it is lexels. However, the metadata for each text at the extremes of the plot can be compared. Figure 7.11 shows the distribution of the top and bottom 10 texts from figure 7.10 according to type and date. The model results in Model 3 showed that burgh records were more likely to contain abbreviated npl and vps {S} tokens than other text types, and state documents less likely. This is indicated in figure 7.11, as the texts which have more abbreviation tokens than predicted include more state documents and fewer burgh records than those texts which include fewer abbreviation tokens than predicted. Similarly, the model indicates that the likelihood of abbreviation increases over time. This is indicated in figure 7.11, as the texts which include fewer abbreviation tokens than predicted are, on average, dated later than those for which the model predicted more tokens than the observed reality.

FIGURE 7.11: Fitted values for the random effect of individual text in a generalised additive model showing the significant effects of SFL, date, location and text type on likelihood of npl and vps {S} morphemes being realised as <ſ> in INFLAOS.



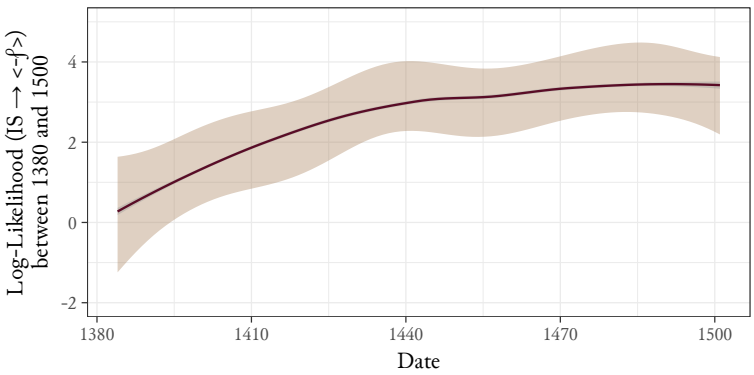
Having said that, these trends relating to type and date in the 20 texts are not particularly strong - in the date plot in figure 7.11, for example, the whiskers of both box plots extend almost to the same range. Likewise, in the type plot, the types which the model indicates are significantly correlated with abbreviation, either positively (burgh records) or negatively (state documents), account for only five out of the 20 texts. The variation in the INFLAOS data which can be attributed to individual texts is less predictable based on qualitative analysis of the texts than the variation due to individual lexels was from a similar analysis of tokens of those lexels. This stands to reason - an outlying lexel is very likely an outlier because of the way it is used by multiple scribes. An outlying text, on the other hand, is necessarily determined by the decisions, practices and, potentially, idiosyncrasies of a single scribe. For example, it is understandable that an early-dated state document such as text 132 (Petition to the king of Scots, 1400, Lanark) would be predicted to have a low likelihood of abbreviation, since state-documents and early dates correlate negatively with the likelihood of abbreviation. In reality, text 132 has nine npl {S} tokens, of which six are abbreviated - more than would be predicted by the model. There are 455 npl {S} tokens attested in INFLAOS which come from state documents dated between 1385 and 1410. Of these, 91 (25%) are abbreviated, and the mean percentage abbreviation for the 11 texts dated 1385 to 1410 is 29%, including text 132 and two other texts with a similar level of abbreviation, 130 and 131.

7.2.1.3 Text date

Figure 7.12 shows the log-likelihood of npl and vps {S} tokens being realised as <ſ> between 1380 and 1500. There is a clear correlation between the lateness of a text's creation and the likelihood of abbreviation in the first half of this period, between approximately 1380 and 1440. After 1440, the likelihood of abbreviation reaches a plateau, and does not increase any further. This plot of log-likelihood represents a statistically robust confirmation of the trend shown in the descriptive plot of the percentage of npl {S} tokens realised as

<ſ> in figure 7.3. The smooth fit line in that plot suggested that the percentage of tokens abbreviated rose from 0% at the very beginning of the period to around 60% by 1430, and stabilised at this figure from 1450 to 1500. In the process of fitting Model 3, the distinction between npl and vps was found to be insignificant as a predictor of <ſ> (that is, a model including the PV of grammel does not explain a significantly greater amount of the deviance observed in the data than a model without it), and is therefore not included in the final model, Model 3. However, the GAM methodology is still subject to the same caveat as a descriptive analysis with regard to token frequency. If there were more attested vps {S} tokens in INFLAOS, the final best-fitting model could potentially be different. Generalised additive modelling allows the most robust analysis of the available data, but it is important to note that the low frequency of vps combined with the much higher frequency of npl tokens could be a reason that the INFLAOS tokens of these two grammels appear indistinguishable.

FIGURE 7.12: Estimated log-likelihood of <ſ> for tokens according to text date.

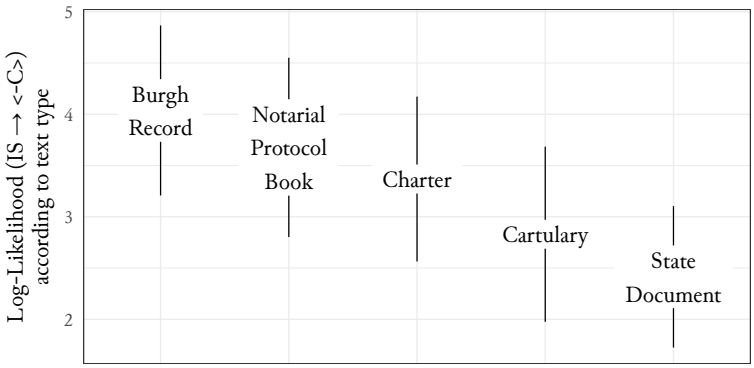


7.2.1.4 Text type

Figure 7.13 shows the log-likelihood of <ſ> estimated by Model 3 for tokens occurring in each text type. There is a clear difference between the likelihood of <ſ> in burgh records at the upper extreme and state documents at the lower extreme. If we accept that abbreviation is inherently a time-saving and convenience-oriented scribal strategy, then it seems intuitively reasonable that it should occur more often in burgh records, which were generally records of proceedings made in real-time, and notarial protocol books which represent the personal record kept by a notary public of his activities. At the other end of this palaeographic convenience spectrum, state documents, by virtue of their association with the court and the fact that they were more likely to be written by scribes with formal training, might be expected to contain fewer abbreviations.

Another factor might be the intended longevity of state documents as opposed to burgh records. Whilst a burgh record might be intended to be referred to at a later date, it is reasonable to assume that a diplomatic

FIGURE 7.13: Estimated log-likelihood of npl and vps {S} to be realised as <ſ> according to text type.



treaty or letter under the privy seal would be expected to be preserved for posterity. This consideration applies, perhaps even more so, to cartularies, which are most often copies[1] of documents deemed important enough to preserve. If a scribe was writing with preservation of the document for future reference in mind, it seems likely that he would prioritise clarity and legibility over speed and convenience. The upper and lower ends of the likelihood scale on which text types are placed in figure 7.13, then, can be seen as opposite ends of a continuum of both textual formality and intention to preserve the contents of a document.

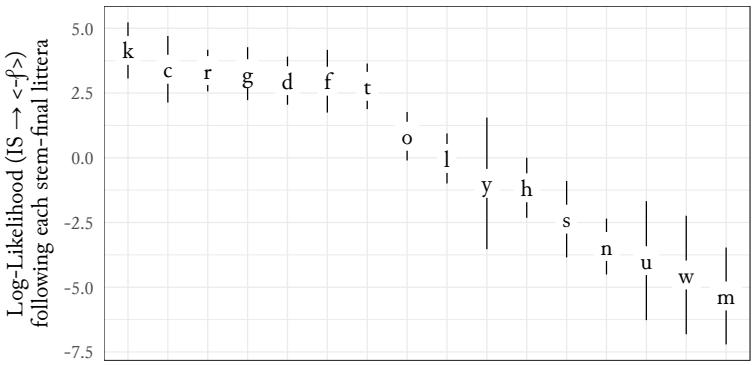
This correlation suggests that {S} abbreviation was increasingly adopted by scribes over time. However, the fact that there appears to be an upper boundary on the likelihood of abbreviation can be explained by the predictability of abbreviation based on SFL.

The likelihood of abbreviation following different SFL is shown in figure 7.14. This plot does not appear to show quite as dramatic a dichotomy in abbreviation following different SFL as the descriptive plot in figure 7.5, which showed that some SFL are followed by <ſ> the vast majority of the time whereas others are very rarely, if ever, followed by <ſ>. The log-likelihood plot instead shows a smoother continuum of SFL, though the difference between those SFL which are most and least likely to be followed by {S} abbreviation is still clear. The reason for this smoother appearance of figure 7.14 is that the GAM model takes into account all of the significant predictors of npl {S} abbreviation shown in the model summary in Model 3. This summary shows that textual metadata (the date, location and type of a text, as well as variation due to individual differences between texts themselves) are significant in predicting whether an npl {S} token will be abbreviated, as is the variation due to individual lexels. The log-likelihood plot in figure 7.14 is a result of all of these significant predictors being taken into account, and the likelihood of a given SFL being followed by <ſ> when all other predictors are held constant. In practical terms, the position of an SFL in figure 7.14 can be interpreted as: the higher the log-likelihood of the SFL, the more

[1] Whilst a cartulary is by definition a collection of copies, there are texts in LAOS which are characterised as cartularies but contain records of legal transactions pertaining to the religious establishment associated with the cartulary. For example, text 1406, from the *Registrum Episcopatus Brechinensis*, is a record of lands leased to the Abbot of Cupar, for which consent was given by the bishop of Brechin.

it is followed by an abbreviation in contexts in which, given the other predictors in the model, abbreviation is less likely to occur, and vice versa. For example, the model shows that tokens from texts with a later date are more likely to be abbreviated, as are tokens from texts identified as burgh records. If stem-final <n> is rarely followed by abbreviation, even in contexts where abbreviation is statistically most likely, then <n> has a low likelihood of being followed by <ſ>.

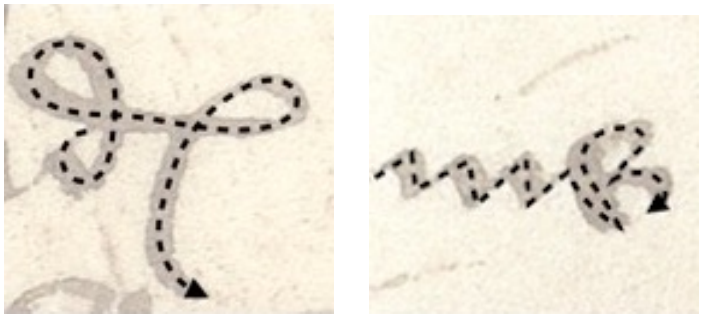
FIGURE 7.14: Estimated log-likelihood of npl and vps {S} to be realised as <ſ> according to stem-final *littera* (SFL).



With regard to the link between the log-likelihood of abbreviation over time as shown in figure 7.12, because {S} abbreviation is extremely predictable based on SFL, with some SFL being very likely and others very unlikely to be followed by <ſ>, the likelihood of an npl {S} token being abbreviated can only climb so high. That is, the likelihood of abbreviation is constrained by the restriction of abbreviation to certain SFL. If some SFL are very rarely followed by abbreviation, and others are very often followed by it, then there is a systematic restriction which prevents abbreviated forms from entirely dominating the realisation of {S}, regardless of how popular the abbreviated form becomes over time.

The reason for this restriction of abbreviation to environments following certain SFL can be found in the respective orthographic forms of these SFL. Figure 7.16 gives an example of each SFL taken from text 36 [1447, CHA, ELO] . Each SFL is shown in word-final position and preceding {S}. The letters which are often followed by <ſ> are those which terminate in a horizontal stroke. These horizontal strokes proceed from different parts of the anatomy of each letter, but all the letters which are commonly followed by <ſ> have this in common. Conversely, all the letters which are seldom followed by <ſ> culminate in a downstroke (<m, n, h>), a curve (<s>) or a hook (<l>). The <ſ> symbol consistently begins with a horizontal stroke before looping backwards into a descender. Figure 7.15 illustrates how easily a letter ending in a horizontal stroke flows into <ſ>, compared with a letter ending in a downstroke.

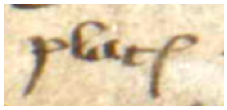
FIGURE 7.15: The path of a pen-stroke for stem-final <d> followed by an abbreviation <ſ> and for stem-final <n> followed by a fully-realised <is> form of npl {S}. Both examples from text 36 [1447, CHA, ELO].



(A) <dſ>

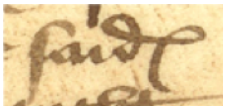
(B) <nis>

FIGURE 7.16: Manuscript examples of each SFL in word-final position and preceding an {S} inflection. SFL ending in a horizontal pen-stroke are generally followed by <ſ>, whereas those ending in a minim stroke are generally followed by a fully-realised form of {S}.

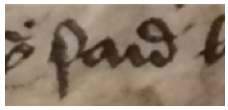


(<C> PL.)
<placſ>
places

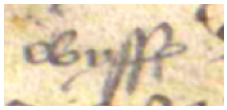
(<C> SG.)
<sic>
such



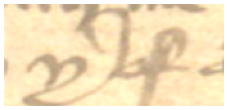
(<D> PL.)
<saidſ>
'[afore-]said [things]'



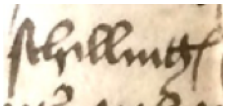
(<D> SG.)
<said>
'[afore-]said [thing]'



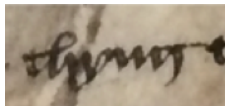
(<F> PL.)
<wyfſ>
wives



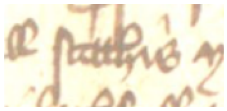
(<F> SG.)
<y(air)of>
thereof



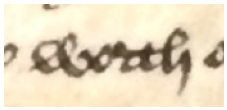
(<G> PL.)
<schillingſ>
shillings



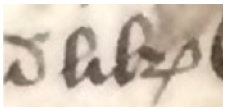
(<G> SG.)
<thyng>
thing



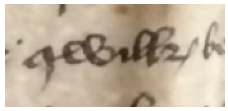
(<H> PL.)
<scathis>
scathes 'damage'



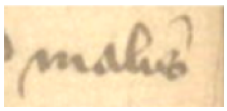
(<H> SG.)
<worth>
worth



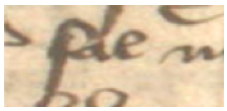
(<K> PL.)
<likſ>
likes



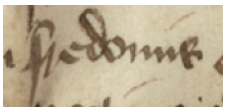
(<K> SG.)
<qwilk>
which



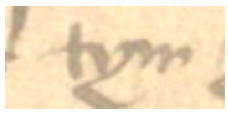
(<L> PL.)
<malis>
mails



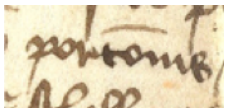
(<L> SG.)
<sal>
shall



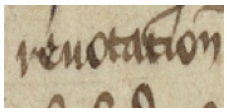
(<M> PL.)
<fredomis>
freedoms



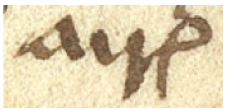
(<M> SG.)
<tym>
time



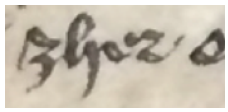
(<N> PL.)
<port(i)onis>
portions



(<N> SG.)
<reuocation>
revocation

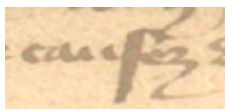


(<R> PL.)
<ayrſ>
heirs

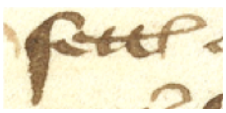


(<R> SG.)
<ȝher>
year

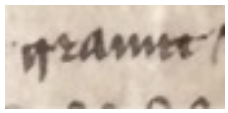
(<S> SG.)
<hors>
horse



(<S> PL.)
<causeȝ>
causes



(<T> PL.)
<settſ>
sets



(<T> SG.)
<grauntt>
<grant>

Chapter 8

Covered Inflectional Vowel Syncope

8.1 Introduction

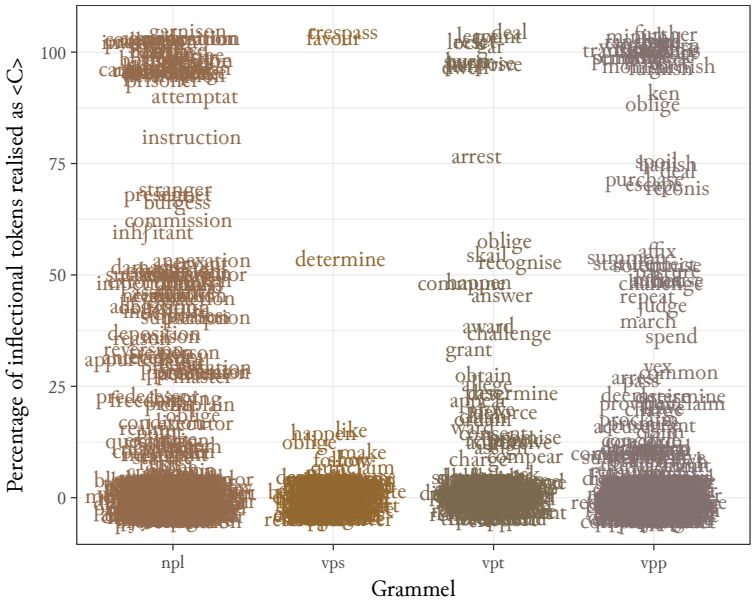
This chapter presents an investigation of the factors correlating with the realisation of plural noun (npl) {S} and past tense verb (vpt) and past participle (vpp) {D} inflections in the form <C>. These forms lack a covered inflectional vowel and are therefore characterised as 'syncopated' inflection forms, though it is acknowledged that orthographic syncope may not correlate with the phonetic loss of inflectional vowels.

Figure 8.1 shows the percentage of *Inflections in A Linguistic Atlas of Older Scots* (INFLAOS) {S} and {D} tokens for each grammel attested as a syncopated form. A 'syncopated form' refers to a consonant-only inflection, with no covered inflectional vowel. Clearly, the occurrence of <C> forms of present tense verb (vps) inflection is very low in INFLAOS. Consequently, this chapter will deal with syncopated forms of npl {S} and vpt and vpp {D} only.

Section 8.2 begins by showing the percentage of npl {S} tokens in each INFLAOS text and lexel which are realised in the form <C>, that is, inflections realised as a consonant only, with no covered vowel. These will be referred to as 'syncopated inflections', referring to their orthographic form. The distribution of percentage values across texts indicates in both cases a general low occurrence of orthographically syncopated inflections, but also shows significant outliers, indicating that a significant amount of syncopated forms are likely to be attributable to scribal preference. There are fewer outliers in the lexel distribution, but combined with the low average percentage of syncopated {S} per lexel, the presence of these outliers is enough to suggest that particular lexels may account for a significant amount of the syncopated forms in INFLAOS. A qualitative examination of the data provides an example where these two factors combine in an idiosyncratic scribal use of syncope with one particular lexel.

To investigate the potential reason for the existence of 'high-syncope' texts, the same percentage-per-text

FIGURE 8.1: Realisation of npl, vps, vpt and vpp inflections as syncopated forms. Only lexels with >10 tokens included.



values are plotted according to type, which reveals that the occurrence of syncopated {S} is higher in burgh records and state documents. These two text types are not intuitively similar, particularly given the conclusions reached in chapter 7 that the orthographic and stylistic choices of high court and burgh court scribes are likely to be quite opposite, as evidenced by the difference in their use of {S} abbreviation. In the case of syncopated {S}, a fundamental difference is shown when tokens in each text type are plotted separately according to their etymology. Whilst the vast majority of the syncopated forms in state documents are non-Germanic, the opposite is the case in burgh records. High court and burgh court scribes clearly use syncope differently to one another.

However, lexel etymology is confounded with stem syllable count - non-Germanic lexels, on average, have a higher number of stem syllables than Germanic lexels. Section 8.2.1 combines all of these predictors in a generalised additive model (GAM) which confirms the significance of text type and etymology as a predictor, even allowing for individual text and lexel variation as random effects. The potential correlation of syncopated {S} forms with syllable count is also considered, following the conclusions reached by Aitken and Macafee (2002) regarding the tendency toward syncope following unstressed syllables ending in a liquid or nasal consonant. Syllable count is found to correlate significantly with {S} syncope as well as the interaction between text type and lexel etymology.

Section 8.3 considers syncopated forms of {D} which, like syncopated {S} forms, appear to be accounted for by particular lexels and texts, in this case, there are a large number of outlying lexels covering the whole percentage scale, suggesting that syncope is confined to these particular lexical contexts.

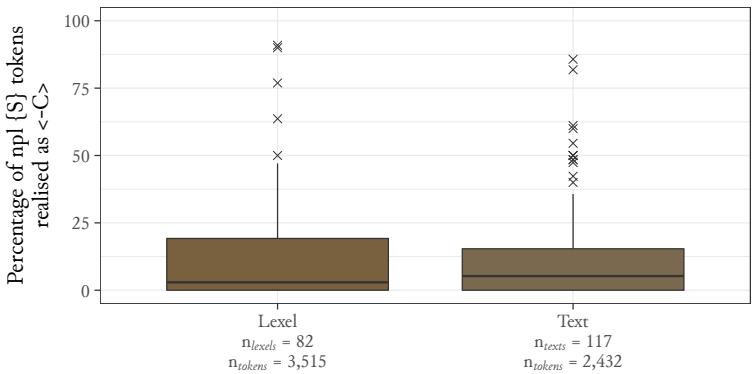
The distributions of syncopated vpt and vpp {D} are then compared, revealing that the texts which are

high in syncopated {D} are likely to contain more vpt than vpp syncope. In particular, burgh records are more likely to contain syncope in vpt than vpp. Two separate GAMs are fitted in section 8.3.1, one for vpt tokens and one for vpp tokens. The conclusion from these two GAMs is that vpp syncope can be fairly well predicted based on various contextual and lexical factors, whereas vpt syncope correlates significantly only with text location. Visualising the likelihood of vpt and vpp syncope using heatmaps shows that the area of highest likelihood is the same for both. This presents the possibility that vpt syncope could itself be influenced by the level of scribal use of vpp syncope.

8.2 Syncopated npl inflections

Figure 8.2 shows the variation in the percentage of npl tokens realised with a syncopated inflection across different texts and across different lexels. A similar graph showing variation in abbreviation percentage was shown in section 7.1, and it was clear that occurrence of abbreviation varied far more widely across different lexels than across different texts. The conclusions of that investigation effectively explained this difference by showing that the use of abbreviation as opposed to full forms by Older Scots (OSc) scribes was strongly conditioned by a single lexical feature - stem-final *littera* (SFL). The variation across texts showed a high mean percentage of abbreviation across texts, with outliers clustered at the very bottom of the scale. This suggests that a high level of abbreviation was the norm for OSc scribes. This same pattern in variation between texts as opposed to lexels is not observable for syncope. Rather, the degree of variation between texts and between lexels appears very similar.

FIGURE 8.2: The variation in the percentage of npl tokens realised with a syncopated inflection across different texts and across different lexels. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown.

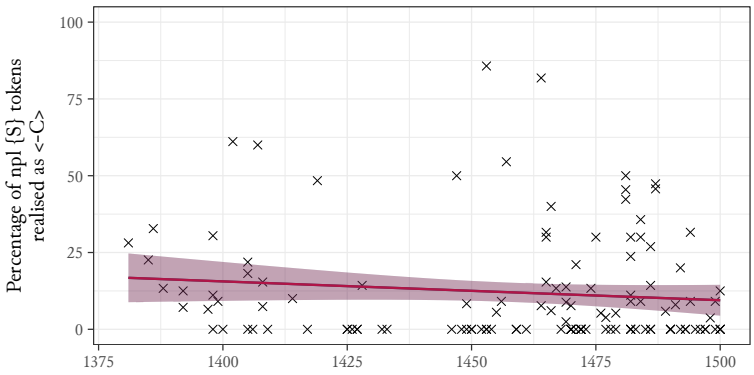


The distribution of this variation in both cases is a low mean percentage of syncope across both texts and lexels, with a cluster of outliers at the higher end of the percentage scale. This suggests that much of the occurrence of syncope in OSc may be attributable to particular texts (and hence particular scribal

habits), and particular lexels which are more likely than others to be syncopated. Unlike the between-lexel variation shown for abbreviation, where an extremely wide interquartile range together with an absence of outliers indicated a stark contrast between SFL usually followed by abbreviation and those rarely followed by abbreviation, the between-lexel variation shown for syncope suggests a very low incidence of syncope for the majority of lexels, with a small subset of lexels accounting for much of the syncope in INFLAOS.

Figure 8.3 shows the percentage of INFLAOS tokens realised with a syncopated inflection over time. Each point on the graph represents an individual text, and Only texts containing more than 10 non-zero, non-abbreviated inflections are shown. Though the points do appear to indicate a higher prevalence of syncope before 1425 and after 1475 than in the intervening period, this is largely an artefact of the paucity of data in the middle of the period (see section 4.1.5.2), as the trend lines in figure 8.3 make clear. The percentage of syncope found in texts does not change in any meaningful way across the period represented by the INFLAOS data. It is clear, however, that, as suggested by the between-text variation in percentage of syncope shown in figure 8.2, there are several texts which represent significant outliers compared to the rest.

FIGURE 8.3: The percentage of npl {S} tokens of each lexel in INFLAOS realised as <C> between 1380 and 1500. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Red: linear trend line; blue: smooth trend line.

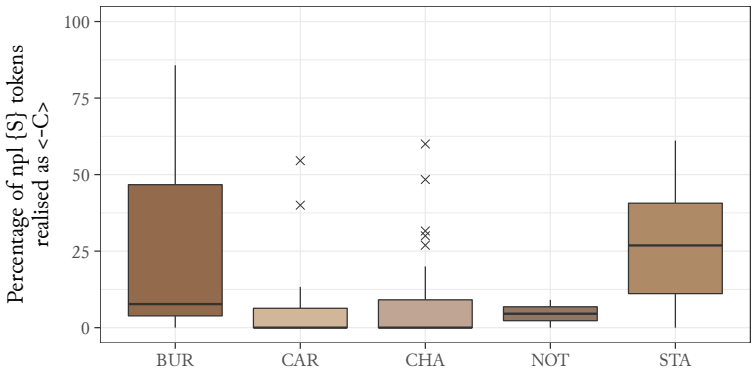


One such text is Text 1116 [BUR, 1453, AYR], an entry from the *Ayr Burgh Court Book*. This text has a total of 14 non-zero, non-abbreviated inflections, of which 12 (85%) are syncopated forms. Notably, of the 12 syncopated forms, nine are instances of the lexel *good*, and the remaining three of *profit*, realised as <gudʒ> and <p(ro)fetʒ> respectively.

Figure 8.4 shows the percentage of INFLAOS npl tokens realised with a syncopated inflection in different text types. There is a large amount of variability between individual burgh record (BUR) and between individual state document (STA) texts. The mean percentage of syncopated forms is however much higher for state documents than for any other text type. This shows that state documents contain, on average, a higher proportion of npl {S} realised as <C> than other text types. The mean percentage of forms realised

as <C> in burgh records is much lower, despite the range of the boxplot for these texts extending higher on the scale than that of state documents.

FIGURE 8.4: The percentage of INFLAOS npl tokens realised with a syncopated inflection in different text types. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown.



This distribution across the percentage scale indicates that the occurrence of syncopated npl {S} in burgh records may be due to a particular group of texts which use them often, but that on average, syncope is not much more frequent in burgh records than in other texts. The group of texts in question is not a small set of outliers, such as those for cartularies and charters - it is extensive enough that the burgh records which have a higher percentage of syncopated forms still fall below the upper quartile boundary. Conversely, the cartularies and charters for which the percentage of syncope is higher than around 25% are classified as outliers. This indicates that syncopated npl {S} is rare enough in these text types that texts which display it, for example, at the average level shown in state documents, are considered unusual.

Figure 8.5 shows the percentage of npl tokens realised with a syncopated inflection according to the etymology of their lexel, with each point in the plot representing a lexel with 10 or more tokens. Overall, non-Germanic lexels have a higher mean percentage of syncopated forms than Germanic lexels, and exhibit more inter-lexel variation.

However, despite this apparent correlation, there are other factors which may confound these preliminary descriptive analyses. It was shown in figure 8.4 that text type appears to correlate with percentage of npl {S} tokens realised as <C>. In particular, state documents displayed the highest overall mean percentage of <C> forms. Figure 8.6 shows the same data as figure 8.4, but with each text type category represented by two boxplots, one including only Germanic, and the other only non-Germanic lexels.

It is clear from figure 8.6 that the correlation between state documents and syncopated forms of npl {S} is dominated by non-Germanic lexels. When only Germanic lexels are considered, state documents do not show a higher mean percentage of syncope than any other text type. Interestingly, when the text type categories are split according to lexel etymology, the mean percentage syncope for non-Germanic lexels is

FIGURE 8.5: The percentage of npl tokens realised with a syncopated inflection according to the lexel etymology. Each point represents an individual lexel. Only lexels exemplified by more than 10 tokens in INFLAOS are shown.

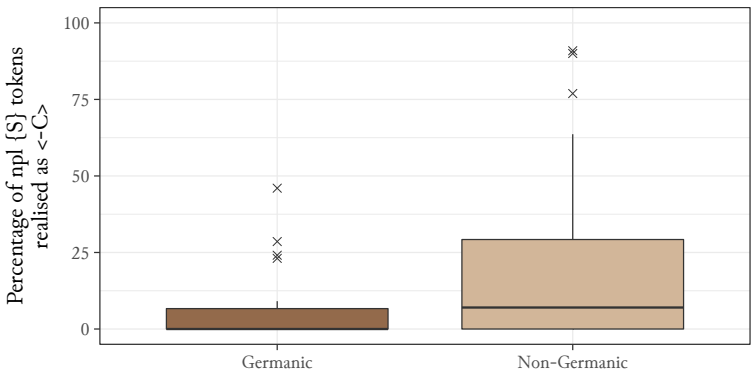
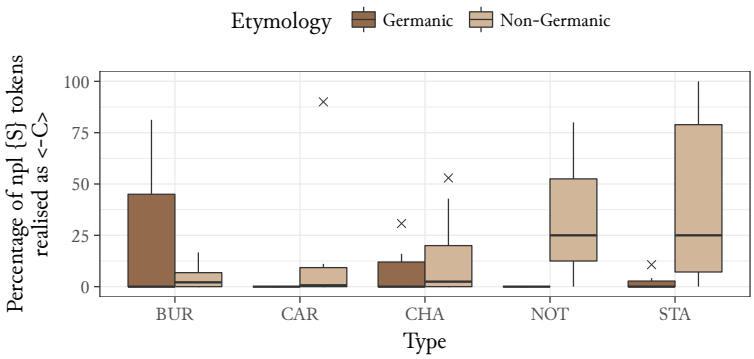


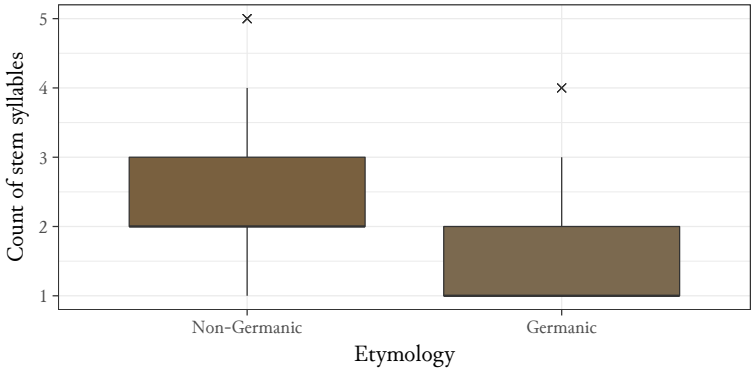
FIGURE 8.6: The percentage of npl tokens realised with a syncopated inflection according to the lexel etymology and text type. Each point represents an individual lexel. Only lexels exemplified by more than 10 tokens in INFLAOS are shown.



the same for state documents and for notarial protocol books. The fact that, in figure 8.4, notarial protocol books were shown to have a far lower percentage of syncopated {S} forms than state documents indicates that scribes of state documents made more use of non-Germanic lexis than those of notarial protocol books: the lower mean percentage of syncope in notarial protocol books is due to the stronger influence on the mean of the greater number of Germanic lexels. It seems that lexel etymology exposes a confounding factor in the apparent correlation between text type and npl {S} syncope. However, figure 8.6 also shows that the correlation between etymology and syncope suggested by figure 8.6 is inadequate when other factors are included in the mix. The higher percentage of syncope for non-Germanic lexels attested in notarial protocol books and state documents contrasts with the lower percentage of these lexels attested with syncopated forms in other text types, suggesting that etymology itself may not be a good predictor of syncope.

In addition to text type, a factor which is likely to covary with etymology is number of stem syllables. Figure 8.7 shows the overall correlation between etymology and stem syllable count for all tokens in INFLAOS regardless of inflection. On average, non-Germanic lexels have more syllables.

FIGURE 8.7: The overall correlation between etymology and stem syllable count for all tokens in INFLAOS regardless of inflection, Each point represents an individual lexel. Only lexels exemplified by more than 10 tokens in INFLAOS are shown.



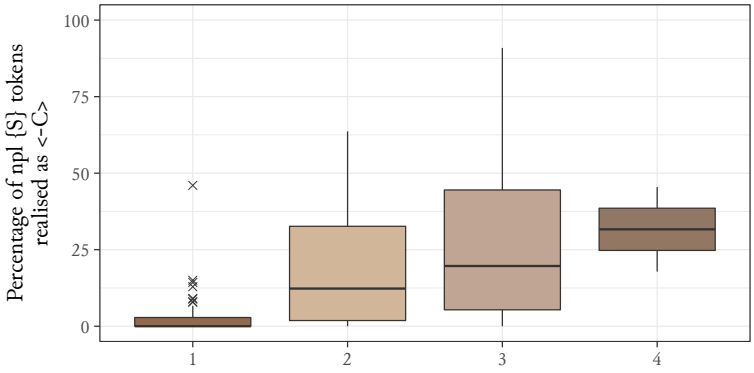
In figure 8.8, which shows the percentage of INFLAOS npl tokens realised with a syncopated inflection according to the syllable count of their stem, *good* again appears as an outlier. The other outlying lexels from figure 8.5 do not: the three lexels which were outliers in terms of non-Germanic etymology, *prisoner*, *attemptat* and *instruction*, are all trisyllabic stems, and fall within the upper quartile boundaries for syncope in this category. Likewise, the other Germanic lexel outliers from figure 8.5, *master*, *freedom* and *bigging*, are disyllabic stems and do not appear as outliers when considered as members of this category. Furthermore, *good* is a much higher outlier than the other outlying monosyllabic-stem lexels, with an approximate average of 30% more syncopated inflection forms than the next highest monosyllabic outlier, *place*. These patterns are likely to be caused by the correlation between stem syllable count and etymology. As shown in section 4.1.2, number of stem syllables is a significant predictor of a lexel's etymology. Specifically, lexels with more stem syllables are more likely to be non-Germanic. The observation that a higher percentage of syncopated inflections occur with nouns with a higher stem syllable count is therefore likely to be related to the observation from figure 8.5 that non-Germanic lexels show a higher overall percentage of syncopated inflection forms.

The general pattern shown in figure 8.8 is an increase in the percentage of npl {S} tokens of a lexel realised as a syncopated form as the number of stem syllables increases. The mean percentage of syncopated npl {S} forms occurring with monosyllabic lexels is practically zero, increasing to an average of around 30% for four-syllable stems. However, this pattern is not necessarily as clear-cut as this distribution makes it appear - figure 8.8 includes only those lexels exemplified by 10 or more tokens.

8.2.1 Modelling the likelihood of covered inflectional vowel (CIV) syncope in npl {S}

Model 4 is a GAM fit to the INFLAOS syncopated npl inflection data. Significant factor levels are indicated with an asterisk. As suggested by the descriptive plots, text type is shown to be a significant predictor of the

FIGURE 8.8: The percentage of INFLAOS npl tokens realised with a syncopated inflection according to stem syllable count. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown.



likelihood that a token will be realised as <C>. Both burgh records and state documents are more likely to contain syncopated realisations of npl {S} than other types. It is interesting to note that both of these types emerge as significant even in the presence of a random effect smooth accounting for individual text variation. As discussed with reference to figure 8.2 (which compared the variation in percentage of syncope between different texts and between different lexels), the overall mean percentage of syncope in burgh records was fairly small, despite the range of values and the interquartile range being large. This suggested a group of texts which use syncopated forms more than the average. Whilst this is still likely to be true, the model shows that burgh records use syncopated npl {S} forms enough that the fact that a text is a burgh record still has predictive power over and above the variation explained by individual texts, as well as date and location.

Stem syllable count is also a significant predictor, with additional syllables in a stem predicting a higher likelihood of syncope.

The smooth functions representing date and location are also shown to be significant, suggesting that use of syncopated npl {S} forms fluctuate significantly over time and space. The significance of these contextual variables, as well as text type as discussed above, exists even in the presence of individual text number modelled as a random effect. Even though there is significant variation in the data which is attributable to the individual differences between texts, the overarching contextual values which group texts together are significantly predictive of the likelihood of an npl {S} token to be realised as <C>.

8.2.1.1 Individual lexels

Figure 8.9 shows a plot of the coefficient estimates of the random effect of lexel in model 4. As suggested by the descriptive exploration of the data in section 8.2.2, the most significant outlying lexel with regard to likelihood of <C> is *good*, which is realised as <gudʒ> in a specific, spatio-temporally localised set of texts.

Figure 8.10 is a map of Scotland showing the county boundaries used in *A Linguistic Atlas of Older Scots*

Model 4: GAM predicting the likelihood of npl {S} to be realised as <C> in INFLAOS according to the significant predictors of text type and location, as well as stem syllable count. The smooths of individual text and lexel capture a significant amount of random variation in the dataset. **R^2 = 0.707; Deviance explained = 70.8%; N = 3,900**

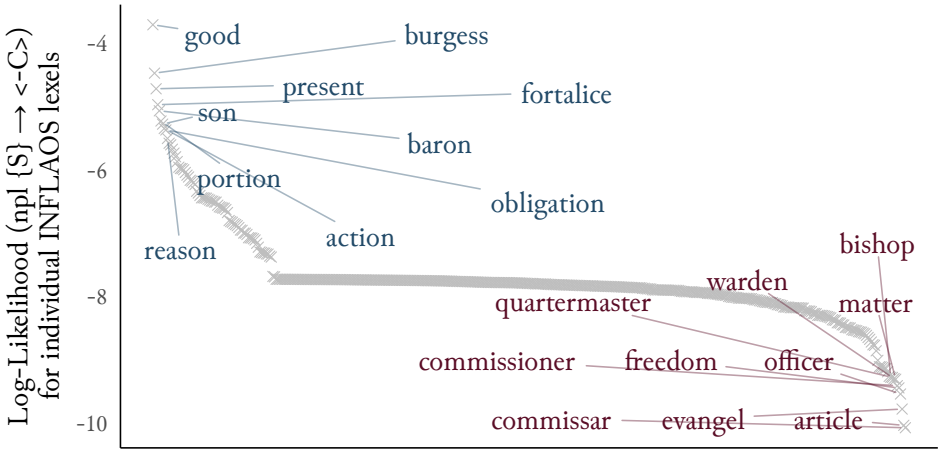
Parametric terms

			Estimate	SE	z	
(Intercept)			-6.01	0.49	-12.19	***
Type	BUR		0.35	0.44	0.79	***
	CAR		0.41	0.48	0.85	
	NOT		1.53	0.49	3.12	
	STA		3.08	0.41	7.59	***
Type * Etymology	BUR	Gmc	1.79	0.74	2.43	*
	CAR	Gmc	0.56	0.88	0.63	
	STA	Gmc	-3.38	1.01	-3.33	***
Syllables	(linear)		2.47	0.36	6.85	***
Etymology	Gmc		-0.34	0.63	-0.54	

Approximate significance of smooth terms:

	EDF	Ref. DF	χ^2	
Text	108.95	699	285.20	***
Lexel	95.19	451	407	***
Latitude, Longitude	7.17	8.81	46.10	***

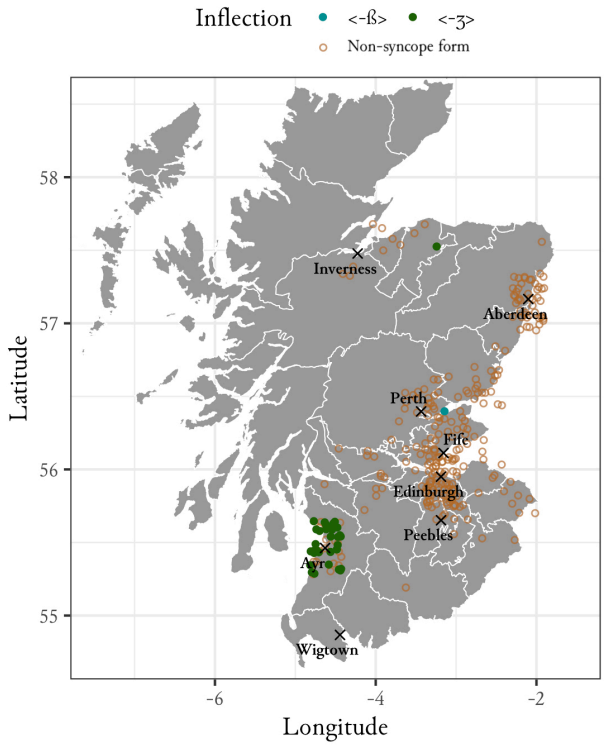
Figure 8.9: Fitted values for the random effect of individual lexel in a generalised additive model showing the significant effects of text type, date and location and stem syllable count on the realisation of npl morphemes as <C> in INFLAOS.



(LAOS). The open points represent all npl tokens of the lexel *good*. The coloured points represent the subset of these tokens which are realised as <C>, with green points denoting the form <ʒ>. It is clear from this visualisation that the realisation of *goods* as <gudʒ> is a phenomenon largely restricted to manuscripts from Ayr and, with the exception of one token in Moray, does not appear elsewhere.

The next highest outliers after *good* are *burgess*, *present* and *fortalice*. These have already been discussed in chapter 6 with regard to their propensity for zero npl inflection. *Present* is ambiguous in the plural

FIGURE 8.10: A map showing the text location of all npl tokens of *good* in INFLAOS. The latitude and logitude values of the points have a small 'jitter' function applied, meaning that they are slightly offset from their actual locations to prevent many points being plotted over the top of one another. This gives a more accurate representation of the number of tokens in a specific location.



form due to its phonetic, orthographic and semantic similarity to *presence*, often realised as <presens>. This ambiguity also exists between other pairs of lexels ending in *-ence* and *-ents*, such as: *absence - absents*, realised as <absens>; and *evidence - evidents*, realised as <evidens>.

In the case of *burgess* and *fortalice*, it was hypothesised in chapter 6, following Kopaczyk (as Bugaj 2004), that the phonetic similarity and potential orthographic similarity between fully-realised {S} and the final syllable of *burgess* caused the reduction of the inflection to zero. In this case, there are two tokens of *burgesses* realised as <burgʒ> in text 251 and one in text 254 (both from the *Registrum de Dunfermelyn* and dated 1457 and 1467 respectively) which, having only one sibilant *littera* representing the second syllable and inflection, appears to be another reduced form. The final <ʒ> is interpreted by Williamson (2008) as an inflection, but there is also evidence in LAOS of scribes spelling singular *burgess* as <burgʒ>, including one instance in a text which is part of the same manuscript, the *Registrum de Dunfermelyn*, as the texts including npl <burgʒ> [text 255: 1472, CAR, FIF].[1] In addition to these three <ʒ> tokens, there is one instance of *burgesses* where {S} is realised as <ß> [text 939: 1493, NOT, PBL]. However, this is an example of *burgens*,

[1]Other texts in which singular *burgess* is realised as <burgʒ>: 1817 (15 tokens), 1819 (1 token), 1828 (2 tokens) and 3012 (2 tokens).

an alternative form from Latin *burgensis* (*Burgen, n.* 2004), as opposed to Old French (OF) *burgeis*, the source of the sibilant-final form (*Burgeis, n.* 2004). The two syncopated forms of *fortalice* (both from text 27, an unlocalised charter dated 1491) are also inflected with <ß> giving the form <fortelesß>. It is less clear where to assign a morpheme boundary for this form than for <burgenß>, which is unambiguously a nasal-final stem followed by a sibilant *littera*. It is possible that the final <ß> is acting as an inflection following the stem-final sibilant, though the vast majority of npl tokens in LAOS ending with <sß> are treated as having zero inflections, as exemplified in (22). This hiving-off approach is supported by the presence of stem-final <sß> in singular nouns, such as the examples in (23).

(22) a. <our luff[t] **burgesß** wilȝa(m) smayl & John̄ morthowsoñ>
*our [be]loved **burgesses**: Wilȝam Smayl and John Morthowson* [text 1531: 1470, BUR, PBL]

b. <al & sindry hyß accionis **causß** & querrellſ>
*all and sundry his actions, **causes** and quarrels* [text 112: 1489, CHA, PTH]

(23) a. <Andrw of cwlañ **burgesß** of Ab(er)deñ geff c(er)tañ a(n)nwallſ til ye kyrk>
*Andrw of Cwlan, **burgess** of Aberdeen, gave certain annuals* [annual payments] *to the ḳirk* [text 1645: 1445, BUR, ABD]

b. <And ȝet ye sad **causß** is co(n)tenvyt but p[e]iudisß of ony p(ar)ty tillꝫ fryday>
and so the [afore]*said **cause*** [court proceedings] *will be continued* [adjourned] *without prejudice of* [disadvantage to] *any party, until Friday* [text 1747: 1457, BUR, ABD]

The examples in (22) and (23) show plural and singular tokens of the nouns *burgess* and *cause*. The specific instances of these tokens have been selected because they unambiguously convey the grammatical number of the lexel in question. (22a) contains the word *burgesses* clearly used to refer to two people, whereas (23a) uses the same form, <burgesß>, to refer to one person. (22b) contains plural *causes* as part of a formulaic list of plural nouns preceded by the adverbial *all*, whereas the same orthographic form, <causß>, occurs in (23b) with a singular verb predicate *is* (used together with *yet* to express the sense of 'will be').

The lexels at the lower end of the likelihood scale in figure 8.9 are those which Model 4 would predict to have a higher likelihood of syncope, based on the other significant predictor variables (PVs), than is actually observed in INFLAOS. All of these lexels are polysyllabic, and the majority end in [r], [n] or [l]: *commissar*, *commissioner*, *officer*, *matter*, *quartermaster*, *warden*, *evangel* and *article*. Aitken and Macafee (2002: 71) discuss the syllabicity, or lack thereof, of {S} in OSc verse. They conclude that in OSc verse, whilst the orthographic realisation of {S} is generally <is> or <ys>, the phonetic realisation is always non-syllabic, "except when the preceding stem syllable consists of a fully unstressed vowel + a liquid or nasal". In this environment, they state, it was possible (though not inevitable) for the unstressed stem syllable to be syncopated in lieu of the inflectional syllable. Examples given by Aitken and Macafee of words where this could

occur include *elders*, *waters* and *nobles*, the phonetic result being [ɛldɹɪs], [waːtɹɪs] and [noːblɪs] respectively (phonetic transcriptions mine). Despite the fact that INFLAOS is a corpus of legal prose rather than verse, it does seem that the effect described by Aitken and Macafee is evidenced by the relative unlikeliness of syncopated {S} forms following some stem-final liquids and nasals. If there is a tendency in the spoken language to preserve the syllabicity of {S} at the expense of the stem-final syllable in such environments, then those environments are presumably the least likely to reflect phonetic syncope in their orthographic realisation.

Having said that, the core assumption upon which the above reasoning is predicated is that orthographic syncope reflects, or is at least influenced by, phonetic syncope. Aitken and Macafee suggest that {S} was inevitably non-syllabic in OSc verse with the potential exception of the environment just described, and the few other studies of OSc inflections have come to a similar conclusion, going further to characterise {S} as non-syllabic in OSc speech in general (Kuipers 1964; Kopaczyk 2001). The tendency in the orthographic practice of OSc scribes would then presumably be towards an orthographic representation of this phonetic syncope, though there is no reason to predict that this would be at an advanced or even intermediate stage in the period covered by LAOS. Indeed, the likelihood of syncope over the period 1385-1500 as estimated by Model 4 and represented in figure 8.3 is fairly static - there appears to be a slight decrease in the likelihood of npl syncope over time, but the wide confidence bands show that, despite the fact that date as a PV explains enough of the variance in the dataset to make it worth including in the final model, it does not reliably show an overall trend in the likelihood of syncope.

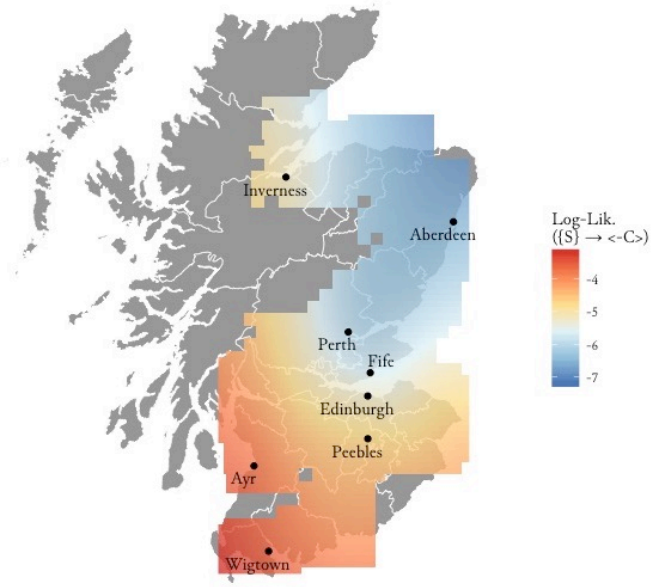
8.2.1.2 Text location

Figure 8.11 is a heat map showing the relative log-likelihood of npl {S} to be realised as <C> in INFLAOS. This realisation is most likely in the South, especially the South-West. It was shown in section 8.2.1.1 that the high percentage of tokens of *goods* with a syncopated {S} form is concentrated in texts from Ayr, but figure 8.11 shows that even taking this into account, syncopated forms are more likely in Ayr and the surrounding areas than in more northerly or central locations.

8.2.1.3 Text type

Figure 8.12 shows the estimated log-likelihood of npl {S} being realised as <C> for each text type. Contrasting this plot with the descriptive plot of text type as a PV in figure 8.4, burgh records clearly stand out much less from charters, cartularies and notarial protocol books in terms of log-likelihood than they did in terms of percentage syncope. This is to be expected, given that the GAM, unlike the percentage plot, takes into account the other factors which shed light on this trend in burgh records. In particular, the random effect of individual lexel is included in the model. figure 8.9 showed that *good* was the most

FIGURE 8.11: A heatmap showing the relative log-likelihood of npl {S} being realised as <C> over the geographical area represented by LAOS.



noticeable outlier with regard to the log-likelihood prediction made by the model. The fact that figure 8.12 shows the likelihood of syncope in burgh records to be more in line with other text types reflects that the apparent correlation of syncope with burgh records was in large part due to tokens of <gudʒ> used in the *Ayr Burgh Court Book* between 1440 and 1460. Having said that, model summary 4 does not suggest that a text being a burgh record has no effect on syncope whatsoever - it is still estimated to be a significant predictor of npl {S} syncope. This shows that, whilst the correlation between burgh records and likelihood of syncope can be explained in large part by the contents of texts from a single manuscript over a 20-year period, Scribes of burgh records were overall more likely to use syncope than those of charters, cartularies and notarial protocol books.

Figure 8.12 shows that the text type most likely to contain syncopated npl {S} is state documents. Figure 8.13 shows the estimated log-likelihood of npl <C> based on the interaction between the PVs of text type and etymology in Model 4. As shown by the coefficient values in the summary table for Model 4, the text types burgh record and state document are both significantly more likely to contain <C> forms of npl {S} than other text types. Both of these types also show a significant interaction with etymology, despite etymology not being significant as a predictor on its own. The positive coefficient value of 1.79 for the combination of burgh records and Germanic etymology indicates that the syncopated tokens in burgh records are more likely to be Germanic, whereas the negative coefficient value of -3.38 for state documents indicates that the syncopated tokens in those texts are more likely to be non-Germanic.

The effect of etymology in conjunction with text type is more pronounced for state documents. In

FIGURE 8.12: Estimated log-likelihood of syncope for tokens according to text type.

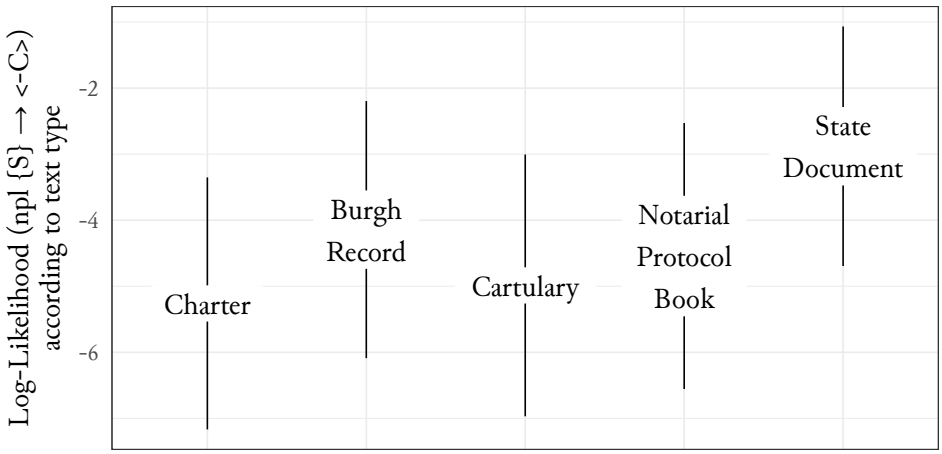
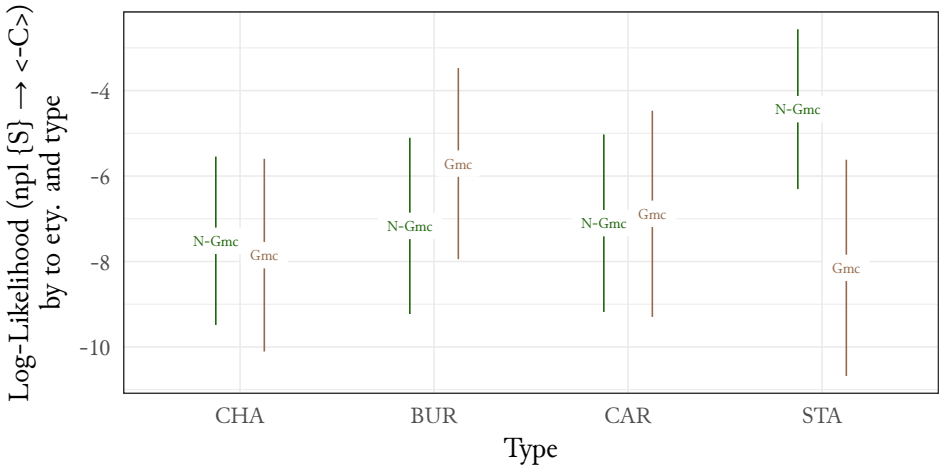


FIGURE 8.13: The estimated log-likelihood of npl <C> based on the interaction between the PVs of text type and etymology in Model 4.



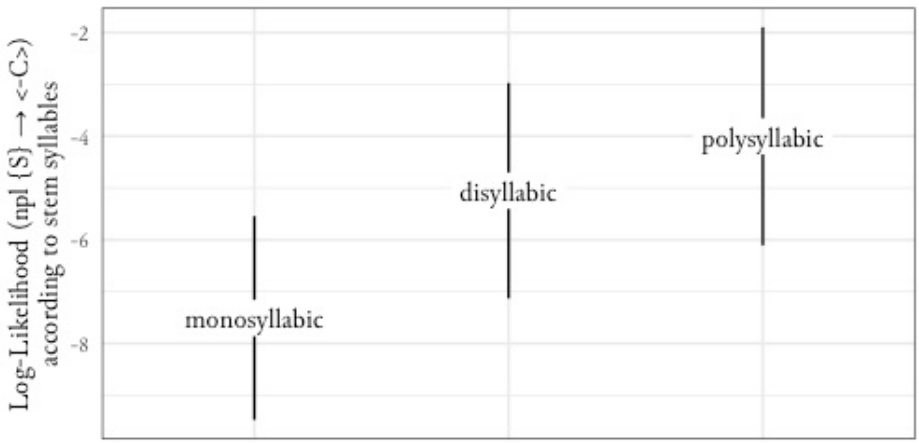
total, there are 212 npl {S} tokens realised as <C> in state documents in INFLAOS, of which 208 (98%) are non-Germanic.

8.2.1.4 Stem syllables

Figure 8.14 shows the predicted log-likelihood of npl {S} being realised as <C> following mono-, di- and polysyllabic stems. The PV of stem syllable count is included in Model 4 as an ordered factor, and stems with three or four syllables are merged into a single category as 'polysyllabic'. This is due to the relatively low number of stems with four syllables, and the fact that, during the initial model fitting process, there was no significant difference identified between three-syllable and four-syllable stems. However, these polysyllabic stems are significantly more likely than disyllabic stems to be followed by <C>, and disyllabic

stems themselves are more likely than monosyllabic stems to exhibit the same inflectional form.

FIGURE 8.14: Estimated log-likelihood of syncope in npl tokens according to stem syllable count.



As described in section 4.1.5.1, the number of syllables in a lexel can be affected by epenthesis. As shown by Model 4, the likelihood of syncope in npl {S} increases with the number of stem syllables, particularly between one and two syllables. It is therefore necessary to consider whether lexels which may be subject to epenthesis are more likely to occur with syncopated forms of {S}. Table 8.1 shows the percentage of npl tokens of monosyllabic noun lexels identified as being susceptible to the insertion of an epenthetic vowel between their two final consonants. The final consonant strings which were used to identify the potential for epenthesis are [lm], [rl], [rm] and [rn]. The lexels identified as containing each of these final strings are listed in table 8.1, along with the number of npl tokens of each lexel occurring with <VC> and <C> realisations of {S}. The top row of table 8.1 groups together all monosyllabic npl lexels which were not identified as being susceptible to epenthesis.

It is clear that, for tokens of the lexels included in table 8.1, the possibility of epenthesis does not correlate with a higher likelihood of syncope. The listed lexels appear no more likely to have a <C> form of {S} than any other monosyllabic lexel, indicating that the orthographic realisation of npl {S} is unaffected by epenthesis.

8.2.2 Inflectional consonants of syncopated npl {S}

Unlike the zero and abbreviated realisations of npl {S} discussed in previous sections, syncopated inflection forms can be categorised according to several distinct forms. The orthographic and morphological homogeneity of the abbreviation symbol <ʃ> was a clear reason to construct a dependent variable (DV) encoding its presence or absence. The current chapter, on the other hand, as with that investigating the zero realisation, deals with a DV which is defined by absence - in this case, the absence of a vowel *littera* between a

TABLE 8.1: The percentage of npl tokens of monosyllabic noun lexels identified as being susceptible to the insertion of an epenthetic vowel between their two final consonants. **PE = potential epenthesis**.

PE	Final string	Lexel	<VC>	<C>	Total
No			2,119 (96%)	87 (4%)	2,206
Yes	**[lm]**	realm	18 (86%)	3 (14%)	21
	[rl]	earl	2 (67%)	1 (33%)	3
	[rm]	arm	1 (100%)	0	1
		ferm	18 (100%)	0	18
		form	1 (100%)	0	1
		harm	8 (100%)	0	8
		storm	1 (100%)	0	1
		term	330 (99%)	3 (1%)	333
	[rn]	bairn	32 (100%)	0	32
		barn	1 (100%)	0	1
		burn	3 (100%)	0	3
		corn	5 (100%)	0	5
		horn	1 (100%)	0	1
Total			**2,540 (96%)**	**94 (4%)**	**2,634**

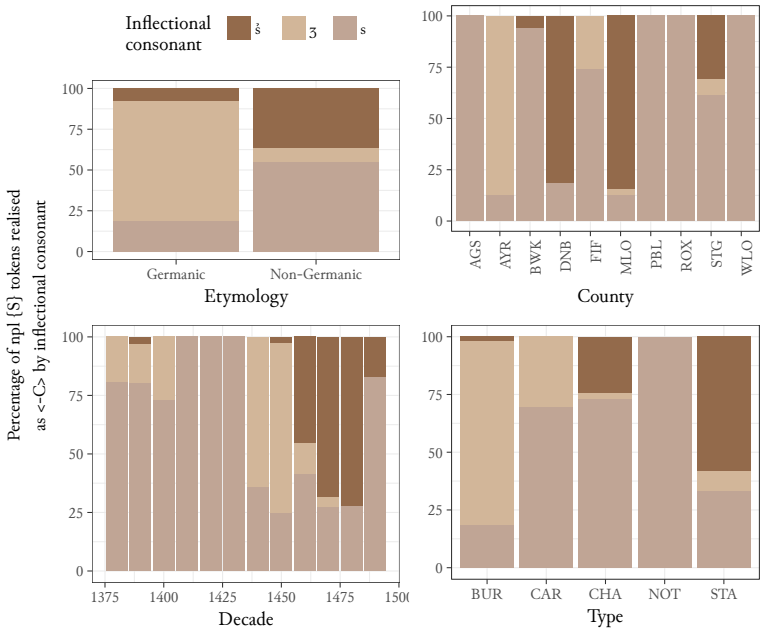
stem-final consonant and an inflectional consonant.

There are three orthographic realisations of npl {S} of the form <C>: <s>, <ṡ> and <ȝ>. Figure 8.15 shows the distribution of these variants according to the etymology of the stem; and the type, date and county of the texts in which they appear. The <s> variant is spread fairly evenly over the different categories of these four PVs, but <ṡ> and <ȝ> appear to be at least somewhat restricted to particular environments.

The summary of model 4 showed that etymology is not a significant predictor of the likelihood of an npl {S} token being realised as <C>. As pointed out in my discussion of the descriptive plot of percentage syncopated forms by stem syllable count (figure 8.8), the correlation of non-Germanic etymology with a higher syllable count likely explains the apparent trend for more syncopated forms in non-Germanic lexels. Although etymology is only a predictor of syncope inasmuch as it reflects syllable count, it might be expected to have some bearing on the consonant used. In particular, the <ȝ> seems intuitively likely to occur more with Romance vocabulary, given the Anglo-Norman origins of the tailed-z *figura*. However, this is not borne out by the etymology graph in figure 8.15 which shows that, of the relatively small number of syncopated Germanic tokens compared with non-Germanic tokens, the majority are <ȝ>. Conversely, only a small proportion of non-Germanic tokens are accounted for by <ȝ>.

Having said that, the other three plots also show clear trends in the use of <ȝ>: this form makes up the vast majority of syncopated forms observed in texts from Ayr, in burgh court books and in the period 1440-1460. Put the other way around: the majority of npl {S} <ȝ> forms come from the *Ayr Burgh Court Book*, 1440-1460. Interestingly, a qualitative examination of these tokens reveals that of 41 tokens of npl {S} realised as <ȝ> in the *Ayr Burgh Court Book*, between 1440 and 1460, 31 (76%) are examples of the

FIGURE 8.15: The distribution of npl {S} <C> realised as <s>, <ŝ> and <ʒ> according to etymology, text type, date and county.



lexel *good*, realised as <gudʒ>. The *Ayr Burgh Court Book* does not contain any examples of npl {S} realised as <s> or <ŝ> within this period, and only two outside it: one instance of <plegs> 'plagues' (LAOS lexel *plage*) [text 1012: 1432, BUR, AYR] and one of <lep(er)s> *lepers* [text 1042, 1436, BUR, AYR].

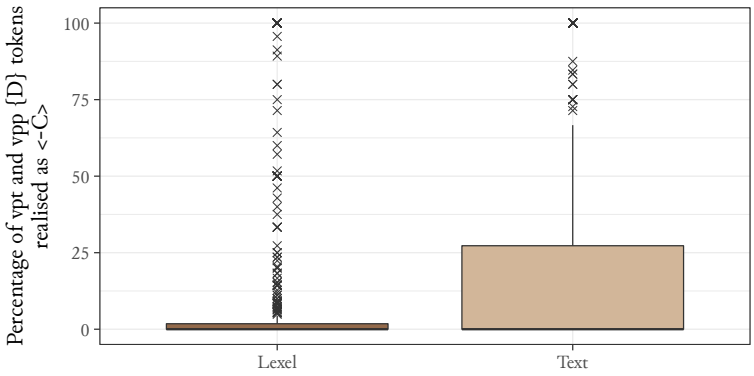
8.3 Syncope in vpt and vpp

Figure 8.16 shows the percentage of vpt and vpp {D} tokens realised as <C> in for each text and for each lexel in INFLAOS. The fact that <C> is a relatively infrequent realisation of vpt and vpp {D} is reflected by the mean percentage scores for both lexels and texts: almost 0% in both cases. The distribution across the percentage scale for individual lexels suggests that whilst for the majority of lexels syncope is used very rarely, if at all, there are a number of lexels which do appear with a syncopated {D} form. This lexel distribution is similar to that observed for npl {S} realised as zero.

The text boxplot in figure 8.16 suggests more systematic inter-textual variation. Whereas any lexel with more than around 5% of tokens syncopated is considered an outlier, texts exhibiting up to 70% syncope still fall below the upper quartile boundary. This indicates that a greater number of texts include syncopated vpt and vpp {D} forms to some degree. Considered together, the distribution of syncopated forms between individual texts and lexels show that syncopated vpt and vpp {D} forms are rare in INFLAOS, but suggests that certain lexels may have a syncopated inflection form continuously across various texts.

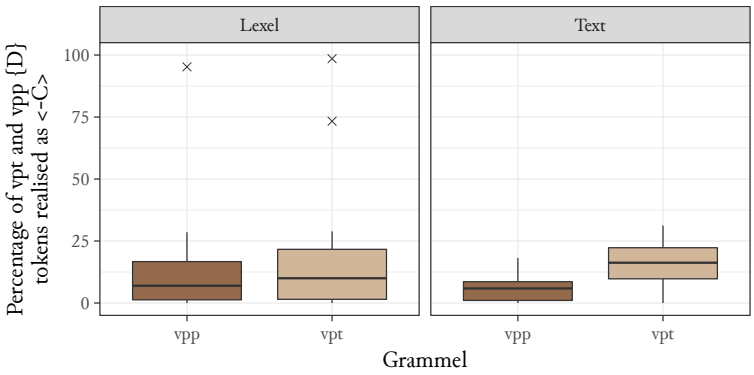
Figure 8.17 shows the percentage of vpt and vpp {D} tokens realised as <C> in INFLAOS. Like fig-

FIGURE 8.16: Percentage of vpt and vpp {D} tokens realised as <C>. The left-hand plot is based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plot is based on the percentage by text (so individual data points represent individual texts). Only texts and lexels with more than 10 tokens are included.



ure 8.16, the two sides of the plot take individual lexels and texts as data points to assess potential differences in the variation between them. In figure 8.17, a separate box plot is also shown for each grammel: vpt or vpp.

FIGURE 8.17: Percentage of vpt and vpp {D} tokens realised as <C>, with vpp and vpt treated separately. The left-hand plots are based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plots are based on the percentage by text (so individual data points represent individual texts). Only texts and lexels with more than 10 tokens are included.



To enable comparison of the distribution of <C> forms in tokens of each grammel, a small subset of INFLAOS tokens was used to compile each half of figure 8.17. The subsets for lexel and text include only data points associated with 10 or more tokens of each grammel. That is, the plot of lexel on the left includes only those lexels with 10 or more vpp tokens and 10 or more vpt tokens; and the text plot on the right includes only those texts with 10 or more tokens of vpp and of vpt.

Figure 8.18 shows the percentage of vpt and vpp {D} tokens of each lexel in INFLAOS realised as <C> between 1380 and 1500. There does not appear to be a significant correlation between date and percentage of <C> for either vpt or vpp tokens. The almost-horizontal shape of both the linear (red) and smooth (blue)

trend lines for vpp look very similar to those in figure 8.3, which showed the same date variable plotted against the percentage of npl {S} realised as <C>. The plot for vpt in figure 8.18 contains no points between 1380 and 1430, as there are no texts from this period containing 10 or more vpt tokens. Even in the period after 1430, points representing texts with 10 or more vpt tokens are much sparser than the corresponding points denoting texts with 10 or more vpp tokens, indicating a higher overall frequency of vpp than vpt forms in the INFLAOS texts.

FIGURE 8.18: The percentage of vpt and vpp {D} tokens of each lexel in INFLAOS realised as <C> between 1380 and 1500. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Red: linear trend line; blue: smooth trend line.

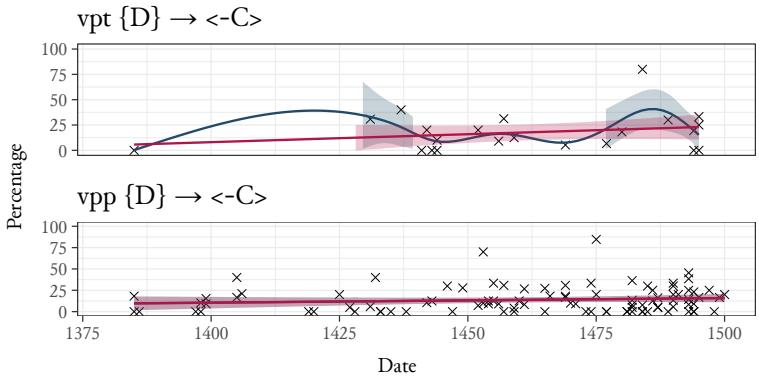
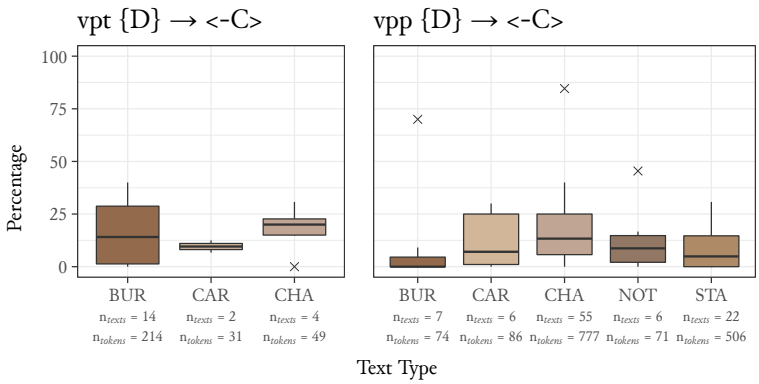


Figure 8.19 shows the percentage of INFLAOS vpt and vpp tokens realised as <C> in different text types. The plot for vpt omits notarial protocol books and state documents, as there is only one of each containing 10 or more vpt tokens. Whilst the paucity of texts containing sufficient vpt tokens is evident here, as it was in figure 8.18, there are nonetheless some apparent correlations shown here.

FIGURE 8.19: The percentage of INFLAOS vpt and vpp tokens in each text realised as <C> for different text types. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Notarial protocol books and state documents are omitted from the plot for vpt as there is only one of each containing 10 or more vpt tokens.



The n-values below the x-axis of the vpt and vpp plots in figure 8.19 show that 214 out of a total 294 (73%) vpt tokens come from burgh records. By contrast, burgh records account for only 74 out of a total 1,514 (5%) vpp tokens.

The mean percentage of vpt {D} realised as <C> in burgh records is higher than the equivalent figure for vpp {D} realised as <C>. This apparent correlation of burgh records with higher incidence of syncope contrasts with the correlation shown in figure 8.4 between state documents and the realisation of npl {S} as <C>.

FIGURE 8.20: The percentage of vpt and vpp {D} tokens realised as <C> for each lexel in INFLAOS, according to stem syllable count. Each point represents an individual lexel. Only lexels exemplified by more than 10 tokens in INFLAOS are shown.

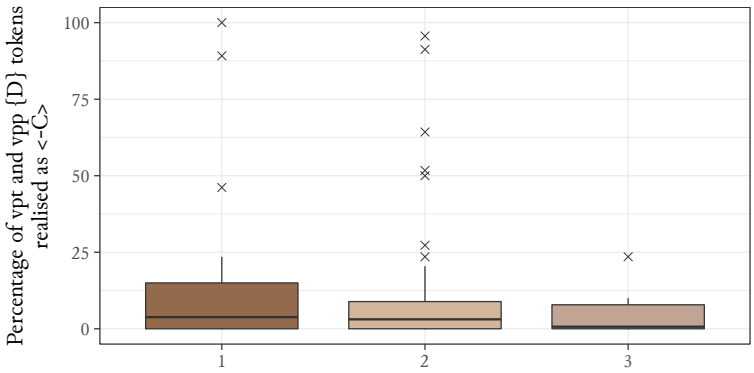


Figure 8.20 shows the percentage of vpt and vpp {D} tokens realised as <C>, according to the number of syllables in a lexel's stem. Though there does not appear to be a very strong correlation between percentage syncope and stem syllable count it is interesting to note that the direction of the correlation is negative. Unlike npl <C> which, as shown in section 8.2, becomes more likely as stem syllable count rises, vpt and vpp <C> appears to be less likely to occur as the stem syllable count rises.

8.3.1 Modelling the likelihood of CIV syncope in vpt and vpp {D}

Model 5 predicts the likelihood of vpp {D} to be realised as <C>. The variables found to be significant predictors of this realisation are text type, SFL and geographic location, as well as the random effects of individual text and lexel. SFL is included in Model 5 in a reduced form, with its categories reduced to <s>, <x> and 'other'. This is because the other SFL categories did not show any predictive significance and were therefore combined into a single category to better show the contrast between them and the predictive SFL, <s> and <x>.

Model 5: GAM predicting the likelihood of vpp {D} to be realised as <C> in INFLAOS according to the significant predictors of text type, SFL and location. The smooths of individual text and lexel capture a significant amount of random variation in the dataset. **R^2 = 0.704; Deviance explained = 69%; N = 4,258**

Parametric coefficients:

		Estimate	SE	*z*	
(Intercept)		-3.34	0.23	-14.35	***
Type	BUR	0.67	0.26	2.58	**
	CAR	-1.33	0.40	-3.35	***
	NOT	-0.38	0.43	-0.87	
	STA	-0.63	0.43	-1.49	
SFL	s	1.51	0.38	4.00	***
	x	3.51	1.04	3.38	***

Approximate significance of smooth terms:

	EDF	Ref. DF	χ^2	
Text	166.50	956.00	356.14	***
Lexel	119.75	380.00	913.44	***
Latitude, Longitude	9.761	12.05	34.56	***

Model 6: GAM predicting the likelihood of vpt {D} to be realised as <C> in INFLAOS according to the single significant predictor of location. The smooths of individual text and lexel capture a significant amount of random var iation in the dataset. **R^2 = 0.566; Deviance explained = 57.1%; N = 1,635**

Parametric coefficients:

	Estimate	SE	*z*	
(Intercept)	-2.80	0.31	-9.04	***

Approximate significance of smooth terms:

	EDF	Ref. DF	χ^2	
Text	69.19	506.00	141.17	***
Lexel	54.72	181.00	276.43	***
Latitude, Longitude	7.768	9.619	75.16	***

8.3.1.1 Individual lexels

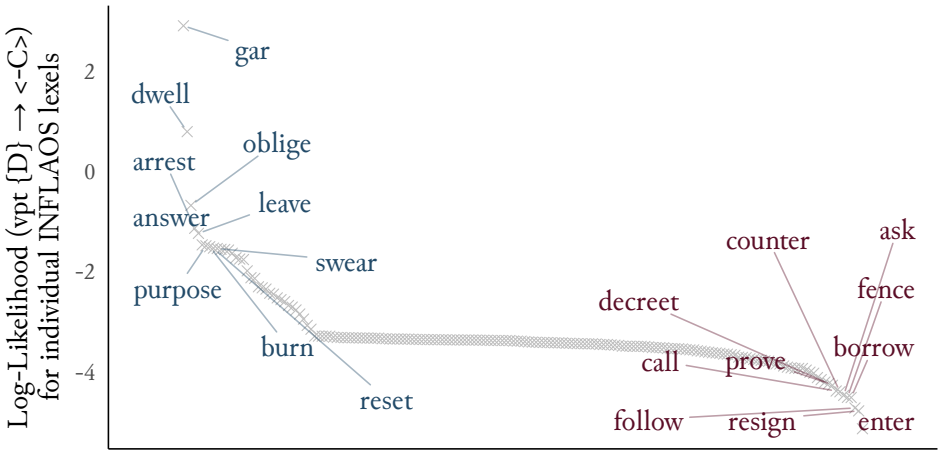
Figure 8.21 shows a plot of the coefficient estimates of the random effect of lexel in Model 6. The lexel with the most 'unexpectedly high' observed likelihood of syncope is *gar*, the vpt form of which is typically realised as <gert> (53 tokens out of a total 64). Aitken and Macafee (2002: 64) note the lowering of [ɛ] to [a] in pre-[r] environments which occurred in Middle English (ME) and OSc, and suggest, based on the orthographic evidence of invariable <e> throughout the fourteenth century, that the earliest date for this change is the late fourteenth century. They note that forms in <a> typically occur later than forms in <e>, a statement which is supported by the LAOS data, in which <a> forms begin to appear only in the second half of the fifteenth century, and then only sporadically. Interestingly, the *Dictionary of the Scots Language* (DSL) entries for *gar* cites vpt and vpp forms with inflectional <it> and <id> as shown in and , whereas the entry for *ger* cites only <gert> as the vpt and vpp form, as shown in (26). This may suggest a change towards the use of the regular {D} inflection after the vowel lowered.

(24) 'Rob Gebson ... **garit** call Rob Boswyll [to] answar to the bowrch'
Rob Gebson called Rob Boswyll to answer to the burgh [court]
(*Burgh Court Book of Dunfermline*, 1497)

(25) 'My siluir pyk tuith quhilk I **garid** Allexander Lowis bring hame out of Flanders'
My silver toothpick which I directed Allexander Lowis to bring home from Flanders
(*Edinburgh Testaments*, 1610)

(26) 'The hail lordis and bordouraris he **gert** bodily be suorne'
All of the lords and borderers [those living in the borders] *he directed to swear a corporal oath*
(*The Acts of the Parliaments of Scotland*, 1448)

FIGURE 8.21: Fitted values for the random effect of individual lexel in a generalised additive model showing the significant effects on the realisation of vpt {D} as <C> in INFLAOS.

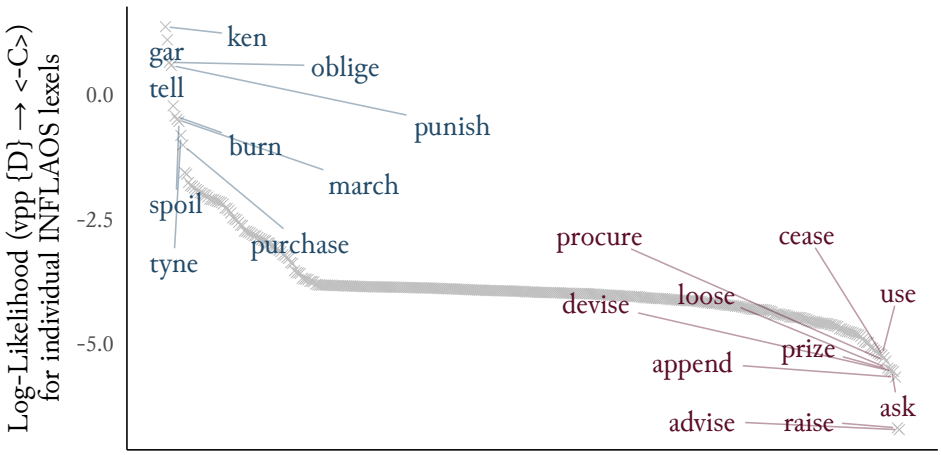


The other noticeable outliers in figure 8.21 are: *dwell*, which has vpt forms in LAOS realised as <duelt> (five tokens) and <dwelt> (one token). There are no instances of *dwelt* with a CIV attested in LAOS, unlike other lexels which show alternation between full and syncopated forms suggesting progressive affixal syncope. This may be due to the etymological root of *dwell*, Old English (OE) *dwellan*, which already had vpt *dwealde* as well as *dwellede* (Bosworth 1898b).

The only other lexel in figure 8.21 which stands out noticeably from the main cluster around the middle of the likelihood scale is *oblige*, which is realised as <oblist> (five tokens), but also with forms containing a CIV, <oblisit>, <oblissit> and <obylyssyt> (one token each). Kopaczyk (as Bugaj 2002: 138) identifies stem-final <s> as an environment in which a "vowel-less dental" vpt or vpp suffix is "phonetically permissible", giving examples of this phenomenon from the *Wigtownshire Burgh Court Book*, including *oblige* realised as <oblist>.

Both *gar* and *oblige* also appear as outliers in figure 8.22, which shows the coefficient estimates of the random effect of lexel in Model 5. However, the highest outlier in this plot is *ken*, which is realised the majority of the time in INFLAOS as <kend> (123 tokens), as in (27), or <kende> (51 tokens), as in (28). Occasionally it occurs with an <it> or <yt> ending (22 tokens), as in (29).

FIGURE 8.22: Fitted values for the random effect of individual lexel in a generalised additive model showing the significant effects on the realisation of vpp {D} as <C> in INFLAOS.



(27) <Be it maid kend till all meń be yir p(rese)nt l(ett)res> [text 4: 1479, n.t., FIF]

(28) <Be it kende til al me(n) be y(ir) p(rese)ntʃ l(ette)ris> [text 11: 1457, CHA ELO]

(29) <Be it kenyt tyl al men thruᵗ yir p(re)sent l(ett)reȝ> [text 172: late 14C, CHA, MLO]

Of a total 203 instances of vpp *ken* in INFLAOS, 175 (83%) form part of the construction *be it* (*made*) *kenned* 'known' *til* 'to' *all men by these present letters* 'writs' [...]. This phrase is a recurring formula which is referred to by Kopaczyk (2013: 246) as a "fixed bundle", a formulaic lexical string which recurred frequently in legal documents of the fifteenth century. Kopaczyk references a humorous rhyme by an OSc clerk which uses this construction, evidencing the status of this phrase as a "characteristic feature of legal jargon" to the point that it could be used as a recognisable basis for parodying the form of legal documents. (27), (28) and (29) show slightly different realisations of this construction (for example: the inclusion of *made* in (27); the use of *through* rather than *by* in (29)), but the formula is still clear in each example.

The formulaic lexical environment in which *ken* occurs may explain the prevalence of the <kend> and <kende> realisations, without which <d> and <de> would be extremely rare forms of vpp {D} in INFLAOS. Of a total 213 <d> and <de> tokens in INFLAOS, 183 (85%) are instances of *ken*. The other instances represent no more than seven tokens of a particular lexel. It may be that the realisation of vpp *ken* takes on this consistent realisation which is not attested with other lexels due to its frequent occurrence within this

formula, and rare occurrence elsewhere. Table 8.2 shows the frequency of vpp *ken* tokens in different text types, indicating whether they occur as part of the *be it (made) kend…* formula. The vast majority of tokens occur in charters (CHA), with some also appearing in cartularies (CAR), which characteristically contained copies of charters. Interestingly, whilst some of the <it> and <yt> forms are non-formulaic instances of *ken*, the highest frequency of <it> and <yt> forms of vpp *ken* occurs as part of the formulaic construction in cartulary copies. There are 10 tokens in this category, each from a different text. Examining the details of these texts individually reveals that all but one are from the same manuscript, the *Copiale Prioratus Sanctiandree*, and dated between 1436 and 1438. It seems plausible that these texts could be the work of a single scribe, or at least scribes who adhered to similar stylistic conventions. Without the original documents which formed the basis of these cartulary copies, it is impossible to say, but these uncharacteristic forms could point to the existence of one or other of what Laing (2004) refers to as "translator" and "mixer" scribes - copyists who 'translated' texts into their own dialect or style, or 'mixed' their own style with that of the original document.

TABLE 8.2: The frequency of vpp *ken* tokens in different text types and whether they occur as part of the *be it (made) kend…* formula.

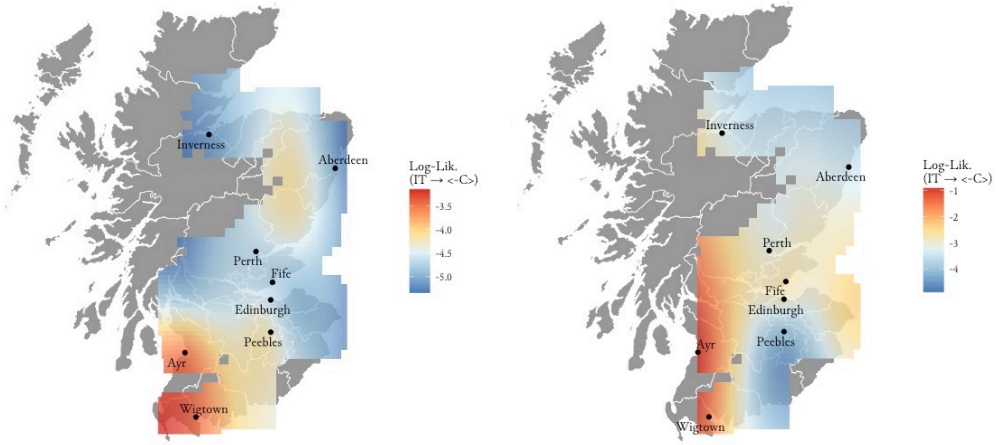
Type	Formula	<d>/<de>	<it>/<yt>	Total
BUR	Yes	1 (100%)	0	1
CAR		31 (76%)	10 (24%)	41
CHA		120 (96%)	5 (4%)	125
NOT		2 (100%)	0	2
STA		0	0	0
	Total (formula = yes)	**154 (91%)**	**15 (9%)**	**169**
BUR	No	0	2 (100%)	2
CAR		0	1 (100%)	1
CHA		17 (100%)	0	17
NOT		1 (100%)	0	1
STA		2 (40%)	3 (60%)	5
	Total (formula = no)	**20 (77%)**	**6 (23%)**	**26**
Total		348 (89%)	42 (11%)	390

8.3.1.2 Text location

Figure 8.23a shows a map of Scotland overlaid with a heatmap indicating the relative likelihood of vpp {D} being realised as <C> over the geographical area represented by LAOS. The area with the highest likelihood of <C> is the south-west.

Figure 8.23b shows the equivalent heatmap for the likelihood of vpt {D} being realised as <C>. The pattern shown in both plots is similar, though figure 8.23b indicates that the likelihood of vpt <C> is more evenly spread and less localised than for vpp. Whereas the vpp heatmap shows that higher likelihood of <C> is concentrated specifically in the south-west counties of Wigtownshire and Ayrshire, the areas of higher

FIGURE 8.23: Heatmaps showing the relative log-likelihood of vpt and vpp {D} being realised as <C> over the geographical area represented by LAOS.



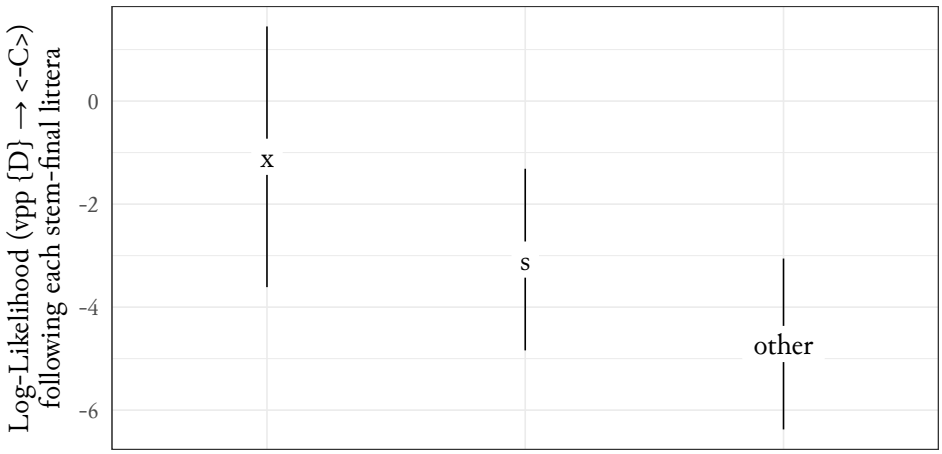
(A) vpp (Model 5) (B) vpt (Model 6)

likelihood of vpt <C> do not stand out as clearly, with higher likelihood towards the western counties, but also, to a lesser extent, in the central area, covering Midlothian and Fife. The pattern of vpt likelihood in figure 8.23b seems to be more generalised than that of vpp in figure 8.23a, with larger areas occupying the middle of the likelihood scale (areas shaded yellow and white).

8.3.1.3 SFL: <s> and <x>

SFL as a predictor is modelled differently to previous models, in that not all *littera* are specified as possible categories of the variable. Instead, the SFL of a token is only labelled specifically if it is <s> or <x>, and all other SFL are subsumed under the category 'other'. 'Other' is treated as the reference level, and therefore is not shown in the summary of Model 5 with a coefficient, standard error or *z*-value. The values shown for <s> and <x> represent the difference in likelihood of a <C> realisation when <s> or <x> occur in stem-final position compared to the likelihood of <C> when any other *littera* occurs in stem-final position. The decision to reduce the dimensions of the SFL PV in this way was motivated by the fact that when a model is fitted including SFL as a PV with all *litterae* represented as category levels, <s> and <x> are the only two *litterae* which show a significantly different likelihood of being followed by <C> from the reference *littera*, which is arbitrarily set as <d>. Put another way, all SFL have a similar likelihood of occurring before <C> to the reference level <d> apart from <s> and <x>, which are significantly more likely than <d> to occur before <C>.

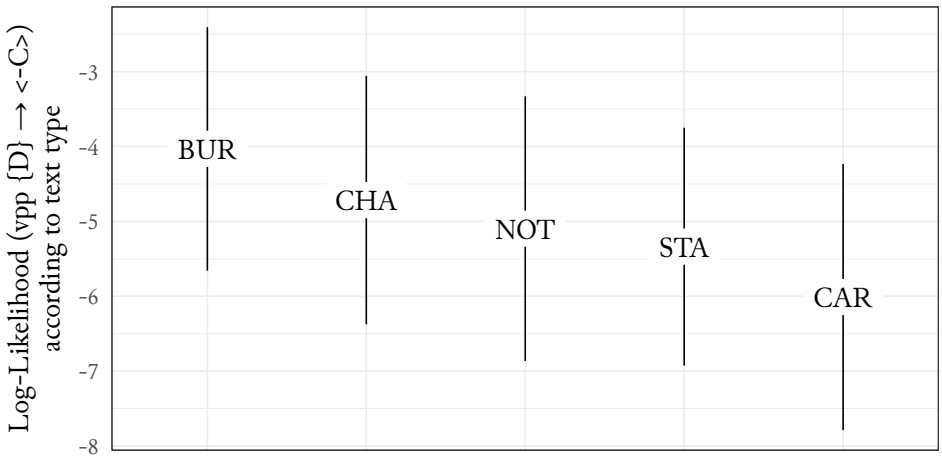
FIGURE 8.24: Estimated log-likelihood of syncope for tokens according to SFL.



8.3.1.4 Text type

Text type is a significant predictor in Model 5, with both burgh records and cartularies showing a significantly different likelihood of <C> to the reference type, charters. The log-likelihood of <C> for each Specifically, burgh records are more likely than charters to contain <C> realisations of vpp {D}, and cartularies less likely. Notarial protocol books and state documents showed no significant difference from charters in this respect.

FIGURE 8.25: Estimated log-likelihood of syncope for tokens according to text type.



8.4 Inflectional consonants of syncopated vpt and vpp {D}

Table 8.3 shows the frequency of occurrence of the inflectional consonants <t>, <d> and written as a superscript to the stem-final *littera* (<^{t}>) in syncopated realisations of vpt and vpp {D}. For both grammels,

<t> accounts for almost half of all inflectional consonants. However, whilst in vpt tokens, the other half of the consonant realisations are <ᵗ>, in vpp tokens, approximately a quarter are realised as <ᵗ> and a quarter as <d>. Only 12 out of a total 225 (5%) vpt syncopated inflections are realised as <d>.

TABLE 8.3: The frequency of occurrence of the inflectional consonants <t>, <d> and <ᵗ> in syncopated realisations of vpt and vpp {D}.

Grammel	**<ᵗ>**	**<d>**	**<t>**	**Total**
vpt	124 (49%)	12 (5%)	115 (46%)	251 (100%)
vpp	172 (23%)	213 (29%)	362 (48%)	747 (100%)
Total	**296 (30%)**	**225 (23%)**	**477 (48%)**	**998 (100%)**

Chapter 9

Variation in the Covered Inflectional Vowel

9.1 Introduction

Table 9.1 summarises the potential realisations of the {S} and {D} covered inflectional vowel (CIV) in *Inflections in A Linguistic Atlas of Older Scots* (INFLAOS). Covered inflectional <i> is by far the most common, accounting for 75% of all {S} and {D} CIV. <y>, generally assumed to form a *litteral* substitution set (see section 1.5.1 with <i> (King 1997: 160; Kopaczyk 2001: 132; Aitken and Macafee 2002: 69; Smith 2012: 29) accounts for 17% of CIV. This ratio of <i> to <y> is similar across all grammels, but covered inflectional <e>, by contrast, has a much higher attestation in plural noun (npl) inflections than in present tense verb (vps), past tense verb (vpt) or past participle (vpp), with 17% of npl {S} tokens having <e> as the CIV.

Table 9.1: The potential realisations of the {S} and {D} CIV in INFLAOS.

Infl.	Grammel	e	i	y	Total
{S}	npl	685 (17%)	2,736 (69%)	570 (14%)	3,991
	vps	25 (2%)	788 (74%)	246 (23%)	1,059
{D}	vpt	4 (0%)	1,216 (85%)	213 (15%)	1,433
	vpp	37 (1%)	3,095 (79%)	786 (20%)	3,918
Total		**751 (7%)**	**7,835 (75%)**	**1,815 (17%)**	**10,401**

This chapter firstly investigates the distribution of covered inflectional <i> and <y>, with the objective of ascertaining whether the occurrence of consecutive minim strokes is the sole conditioning factor in a scribe's use of <y> over the more usual <i>. Secondly, the distribution of covered inflectional <e> is considered.

Section 9.2 reports the results of an investigation into the significance of this effect when considered alongside contextual factors such as date and location, and lexical factors such as etymology and grammel.

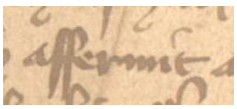
The effects of minim frequency and proximity are investigated to determine the necessary orthographic conditions for substitution of <i> with <y>. For example, do scribes resort to <y> in inflections only when the stem ends with a series of minims, as in example (30a), or can a series of minims earlier in the word, as in (30b), trigger the use of <y>? Alternatively, might scribes resort to other methods of avoiding minim-related ambiguity, such as the superscript <t> realisation shown in (30c)?

(30) a. <sommys> *soams* npl [text 341: 1459, CAR, AGS]

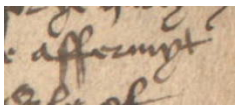
b. <iugyt> *judge* vpp [text 1631: 1444, BUR, ABD]

c. <aff(er)mm^{t}> *affirm* vpp [text 1285: 1468, BUR, FIF]

FIGURE 9.1: Two examples of *affirmed*: in text 1601 with vpt {D} realised as <it> following stem-final <m>; and in text 1614 with vpp {D} realised as <yt> following stem-final <m>.



(A) <affermit> *affirmed* (vpt)
text 1601 [1434, BUR, ABD]



(B) <affermyt> *affirmed* (vpp)
text 1614 [1434, BUR, ABD]

Section 9.4 firstly addresses the issue of ambiguity in hiving off inflections with covered <e> as opposed to stem-final <e> by investigating subsets of both singular and plural nouns. Potential reasons for the presence or absence of final <e> in singular nouns are considered in section 9.4.1, and generalised additive models (GAMs) are fitted to the singular and plural data subsets, showing that the same predictor variables (PVs) are significant for both types of <e>, but that the corresponding trends are different.

9.2 Covered inflectional <i> and <y> in npl {S} inflections

It was established in chapter 7 that the realisation of {S} as the abbreviation symbol <ſ> was conditioned primarily by a specific *figural* feature of the stem-final *littera* (SFL), but was also significantly predicted by contextual variables such as date and location. In this section, I examine another inflectional realisation which appears to reflect *litteral* substitution, namely the realisation of the CIV as <y> as opposed to the more common <i>.

Figure 9.2 shows the percentage of inflectional tokens with covered inflectional <yC> per lexel for each grammel. The distribution of the percentage of <yC> inflections used is fairly even across the different lexels, though vpt {D} inflections have fewer lexels clustered around 100%. Notably, similar plots of the distribution by lexel and grammel for the other dependent variables (DVs) discussed in previous sections have not shown a noticeable disparity in distribution between vpt and vpp. When it comes to the use of <yC> over <iC>, however, vpt {D} appears to show a lower overall incidence of <yC>.

FIGURE 9.2: The realisation of npl, vps, vpt and vpp inflections as <yC> forms as opposed to <iC> forms. Only lexels with >10 tokens included.

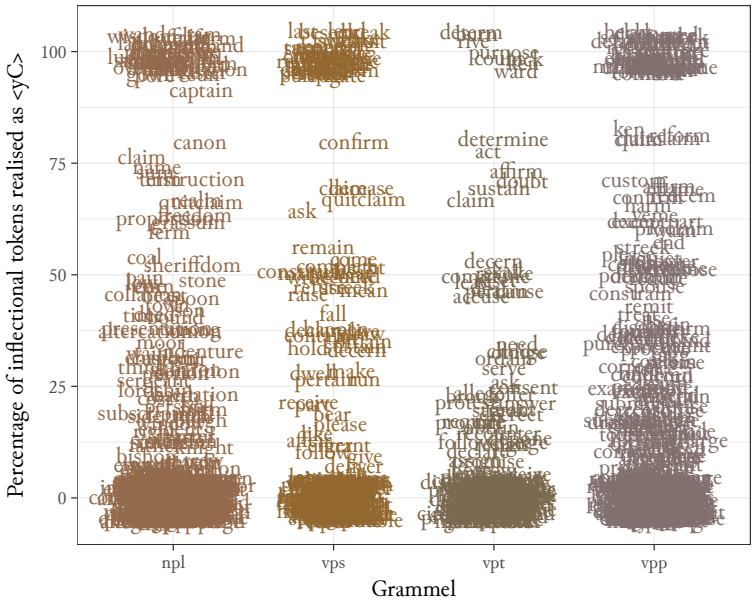


Figure 9.3 shows the percentage of npl {S} tokens realised as <yC>. The left-hand plot is based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plot is based on the percentage by text. The two boxplots are very similar in terms of their range, mean and outliers, suggesting that the degree of <iC>/ <yC> variation between different texts and between different lexels is similar and, therefore, there is unlikely to be a textual or a lexical PV which accounts for a significant proportion of the variation.

FIGURE 9.3: The percentage of npl {S} tokens realised as <yC>. The left-hand plot is based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plot is based on the percentage by text (so individual data points represent individual texts). Only texts and lexels with more than 10 tokens are included.

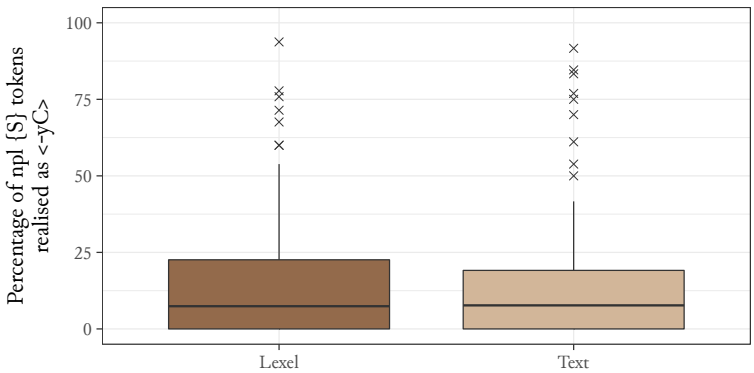


Figure 9.4 shows the percentage of npl {S} tokens in each INFLAOS text realised as <yC> as opposed

to <iC>, arranged in text date order. There is a potential correlation between text date and percentage <yC>, but the confidence band around the trend line fit through the data points suggests that this apparent correlation could be an artefact of the distribution of texts over time in *A Linguistic Atlas of Older Scots* (LAOS).

FIGURE 9.4: The percentage of npl {S} tokens of each lexel in INFLAOS realised as <yC> as opposed to <iC> between 1380 and 1500. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Red: linear trend line; blue: smooth trend line.

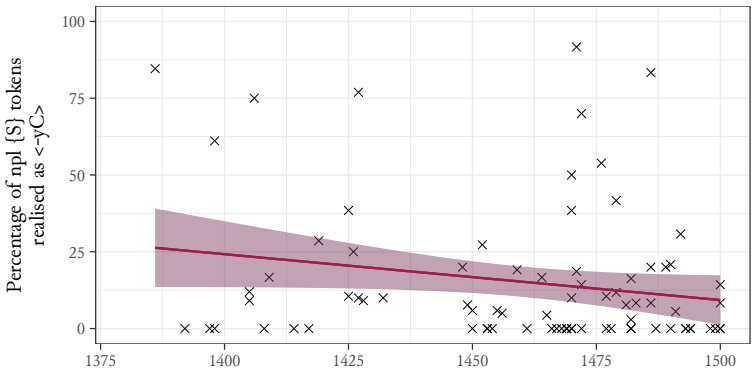
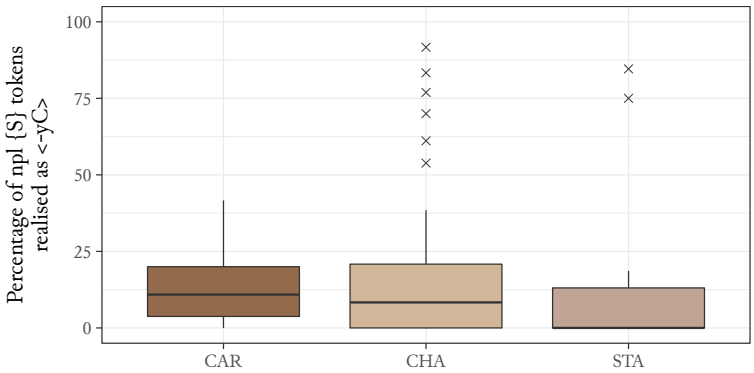


Figure 9.5 shows the percentage of npl {S} tokens attested in each text type realised as <yC> rather than <iC>. Burgh records and notarial protocol books are omitted as there is only one burgh record with more than 10 tokens of npl {S} realised as <iC> or <yC>, and no notarial protocol books. This is due to the high level of abbreviation and corresponding small amount of fully-realised inflections in both of these text types (see section 3.1.3).

FIGURE 9.5: Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown.



The mean percentage of tokens realised as <yC> is around 10% for cartularies and charters, but close to 0% for state documents. There are two noticeable outlying state documents: text 63 (11 tokens out of 13 [75%] realised as <yC>); and text 131 (9 tokens out of 12 [75%] realised as <yC>). Interestingly, in both of

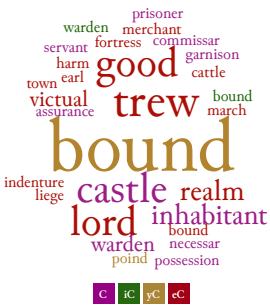
these texts, the tokens of <yC> are dominated by one particular lexel. In text 63, 10 of the 11 appearances of <yC> occur in the word *bounds*, and in text 131, eight of the nine <yC> tokens occur in the word *trews*. *Trews* is the plural form of the noun *trew* < Old English (OE) *trēow* 'good faith', the source of Modern Scots (ModSc) and Present Day English (PDE) 'truce' (Oxford English Dictionary 2018). Both *truce* and *bounds* are words which were most commonly used in the plural rather than the singular. In both cases, the rarity of the singular form of the word is indicated by attestations of the plural form in {S} in syntactically singular environments. There are no examples of this usage in LAOS, but examples (31) and (32) show examples from Older Scots (OSc) texts of *trews* and *bounds* respectively with singular determiners.

(31) <The ambassodouris of France and Ingland came to begg of this King a trews>
the ambassadors of France and England came to beg a truce of this King
(Wyntoun ([1350-1420] 1872-79); cited in *Dictionary of the Scots Language* (DSL) (2004))

(32) <He compasit ane certane boundis>
he encircled a certain bounds
(Livy ([1533] 1901); cited in DSL (2004))

The usage of <yC> in texts 63 and 131 is unusual for OSc state documents, as shown by figure 9.5. Added to this, the use of <yC> by the scribes of these documents represents, in both cases, the use of an orthographic form of npl {S} with a single word, which the scribes do not use elsewhere.

FIGURE 9.6: npl lexels in texts 63 and 131. Word size indicates frequency and colour indicates orthographic realisation of npl inflection.



(A) text 63 [1386, STA, n.l.]

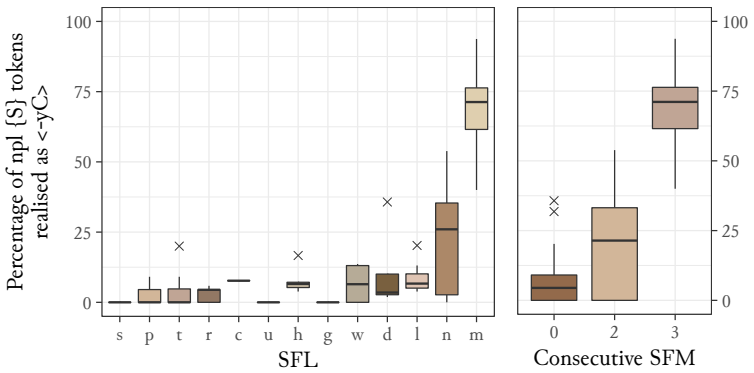
(B) text 131 [1405, STA, DNB]

Figure 9.6 illustrates the distribution of npl tokens in Text 63 and Text 131. Text 63 contains mainly <C> and <eC> realisations of npl {S}, whereas the dominant realisation in Text 131 is the abbreviation symbol <ſ>. The use of <yC> in Text 63 is particularly interesting here, as this text is described by Williamson (2008) as being "in an English hand", and consistently makes use of the characteristically Middle English

(ME) <eC> realisation, as well as <C>, with very little use of the typically OSc <iC>/<yC> forms aside from the tokens of *bounds*.

Figure 9.7 plots the percentage of npl tokens of each lexel in INFLAOS where {S} is realised as <yC> rather than <iC>. Each data point in the plot represents a lexel, and only lexels represented by 10 or more tokens in INFLAOS are included. The lexels are grouped according to token characteristics: in the left pane, lexels have been grouped according to SFL; and in the right pane, according to the number of consecutive stem-final minims (SFMs).

FIGURE 9.7: Each point represents an individual lexel. Only lexels exemplified by more than 10 tokens in INFLAOS are shown.



Some examples of the number of stem-final minims particular orthographic forms are judged to contain are given in table 9.2.

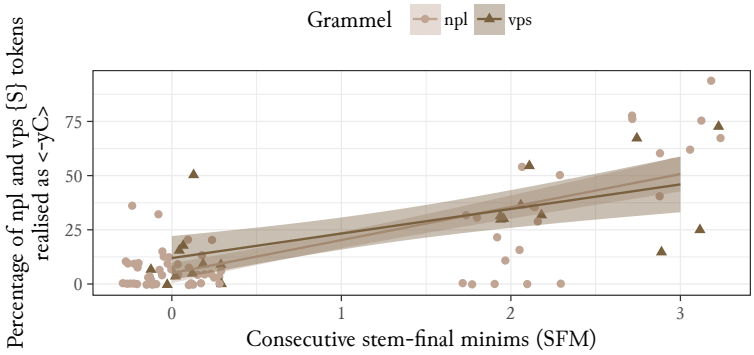
TABLE 9.2: Examples of the number of stem-final minims (SFMs) particular orthographic forms are judged to contain.

Lexel	**Stem form(s)**	**Consecutive SFM**	
liege	*leg-; lieg-*	*-g-*	0
scot	*scott-*	*-t-*	0
brief	*breu-*	*-u-*	2
warden	*wardan-; warden-*	*-n-*	2
fine	*fin-*	*-in-*	3
term	*t(er)m-; ter(m)m-; term-; trem-*	*-m-*	3
custom	*custun-*	*-un-*	4
sum	*su(m)m-*	*-u(m)m-*	5
term	*termm-, tremm-*	*-mm-*	6

9.2.1 Modelling the likelihood of covered inflectional <y> in npl {S}

Model 7 predicts the realisation of npl {S} as <yC> as opposed to <iC> in INFLAOS. The only significant fixed effect in the model is consecutive SFM count. With just this PV together with the significant random effects of individual text and individual lexel, the model explains 84% of the deviance in the dataset (R^2 =

FIGURE 9.8: The correlation between SFM count and the percentage of npl and vps tokens realised as <yC>. Each point represents an individual lexel. Only lexels exemplified by more than 10 tokens in INFLAOS are shown.



0.672). This indicates that the model is a reasonably good fit to the data, and that there is some evidence that scribal choice of <yC> over <iC> can be predicted by the necessity of discriminating letters. However, it is likely that there are other sources of variation which are unaccounted for by the model, or that random scribal choice influences the use of <y> over <i>.

MODEL 7: GAM of the likelihood of npl {S} to be realised as <yC> as opposed to <iC>, predicted by number of consecutive stem-final minims (SFM) and including random effects of individual text and lexel. **R^2 = 0.672; Deviance explained = 65.5%; N = 11,896**

Parametric terms

		Estimate	SE	*z*	
(Intercept)		-3.22	0.18	-18.28	***
SFM	2	1.18	0.24	7.29	***
SFM	3+	4.32	0.31	14.03	***

Approximate significance of smooth terms:

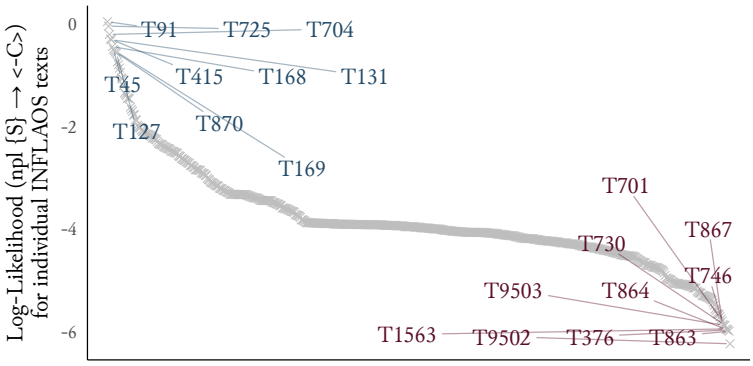
	edf	Ref.df	χ^2	
Lexel	46.42	411.00	128.40	***
Text	221.62	623.00	629.00	***

Consecutive SFM count is coded in Model 7 as an ordered factor. This means that, whilst ordinarily the significance of factor levels in a model is estimated with regard to difference from the reference level, in this case, the significance of each level is estimated with regard to the preceding level. The reference level for SFM is zero, so the model shows that stems ending in two minim strokes are significantly more likely to be followed by covered inflectional <y>, and that stems ending in three or more minim strokes are significantly more likely to be followed by <y> than those ending with two minim strokes.

Figure 9.9 shows the log-likelihood of npl <yC> for individual texts as estimated by Model 7. The texts are arranged along the x-axis in descending order of log-likelihood. The 10 text numbers labelled in blue therefore represent the texts which the model suggests contain more covered inflectional <y> than they should, considering what would be predicted based on the other significant PVs in the model. Text

131, which was discussed in reference to figure 9.5 as an outlier for the category of state documents, is unsurprisingly found amongst the outliers here too. The most significant outlier shown in figure 9.9, however, is text 91 [1385, CHA, PTH]. This text contains seven npl <yC> tokens: three <landys> *lands*; two <endent(ur)ys> *indentures*; one <handys> *hands*; and one <trowthys> *truths*. Given the clear importance of the effect of SFM on the realisation of npl {S} as <yC>, it seems very likely that the reason for text 91's outlier status is the scribe's use of <ys> inflection forms following stem-final <d>. As noted in chapter 7, the <d> *figura* typically finishes with a looped ascender, extending horizontally to the right, as shown in figures 7.16<d> sg. and 7.16<d> pl..

FIGURE 9.9: The log-likelihood of npl <yC> for individual texts as estimated by Model 7.

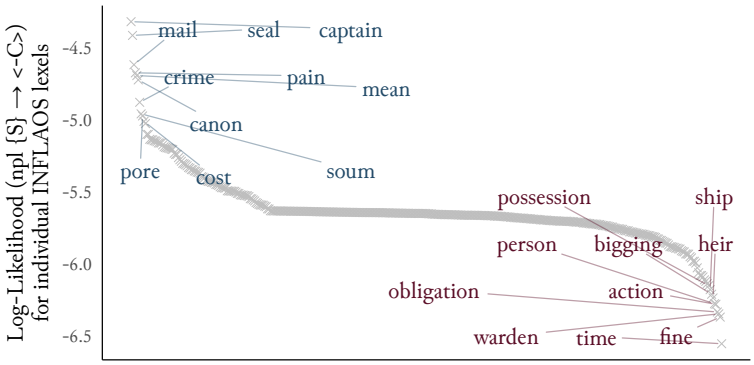


Some of the highest outliers for the random effect of individual text contain tokens with stem-final <l>. This *littera* has previously been discussed in reference to the DV of abbreviation, for which it was shown to be a somewhat ambiguous SFL category, whereas other SFL emerged as either strongly predictive of <ſ> or strongly predictive of its absence. In chapter 7, it was suggested that this predictive ambiguity might be explained by different potential *figural* representations of <l> - either with a looped ascender like <d>, or with a straight ascender. The latter representation would be very much like a vertically extended minim-stroke, which might explain the presence of stem-final <l> tokens in the outlying texts according to Model 7. Using this reasoning, it may even be the case that certain *figural* renderings of <d> with a straight ascender rather than a loop could give rise to the ambiguity-resolving covered inflectional <y>.

Figure 9.10 plots the fitted values for the random effect of individual lexel in Model 7. Lexels are placed on the x-axis in descending order of y-axis value (log-likelihood of <y> as opposed to <i>), therefore lexels closer to the top left corner of the plot are those which have a higher likelihood of occurring with a <yC> inflection that the model would have predicted based on the other factors which correlate significantly with <yC>.

The highest outlier in figure 9.10 is *captain*. There are eight tokens of the npl form *captains* in INFLAOS: all realised as <capitanys>; and all from text 9519 (act of parliament, 1482). Given this fact, it

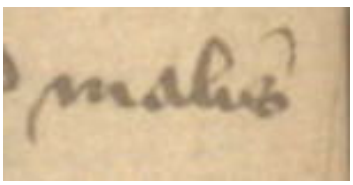
FIGURE 9.10: Fitted values for the random effect of individual lexel in a generalised additive model showing the significant effects on the realisation of npl {S} as <yC> in INFLAOS.



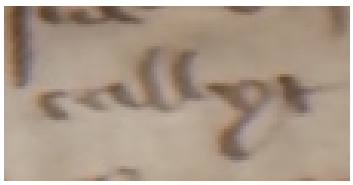
might be expected that these tokens should be attributed to textual rather than lexical variation, but an examination of all npl the tokens occurring in text 9519 reveals that the scribe does not use <yC> for any other plural noun, including three tokens of <wardanis> *wardens*, which also have two consecutive stem-final minims in the same SFL: <n>.

The next highest outliers are *mail* and *seal*. It has been noted previously that stem-final <l> can take more than one form, and that this might account for the 'middling' prevalence of <ſ> realisations of npl {S} following stem-final <l>. The inclusion of individual SFL as a factor in a GAM predicting likelihood of npl <yC> does not suggest a significantly higher likelihood of <yC> over <iC> following stem-final <l>. However, it is possible that the variation in realisation of the <l> *figura* by different scribes has an influence on the following CIV, which only appears in the model as an indication of higher likelihood in these two common <l>-final stems. Figure 9.11 shows two manuscript examples of stem-final <l> followed by a CIV. In 9.11a, the <l> has a looped ascender and is followed by covered inflectional <i>, whereas in 9.11b, the <l> does not feature a loop, which makes its form more similar to the single stroke minim. It is possible that where stem-final <l> was *figurally* less distinguishable from a minim stroke, it was more likely to be followed by <yC>. However, much more through palaeographic research would be needed to assess this.

FIGURE 9.11: Two examples of stem-final <l>: from text 1611 followed by covered inflectional <i> in <malis> *mails*; and from text 4 followed by covered inflectional <y> in <callyt> *called*.



(A) <malis> *mails* (npl); text 1611



(B) <callyt> *called* (vpp); text 4

At the other end of the scale in figure 9.10 are those lexels which, based on the other significant

predictors in Model 7, would be expected to have a higher likelihood of <yC> than the observed data shows. The lowest outlier here is *time* which often occurs with the stem form <tim> and consequently has four consecutive SFMs. The lexel *fine* also often occurs with our consecutive SFMs, and is likewise a low outlier here. This observation may suggest that *time* and *fine* are predicted too high a likelihood of <yC> on the grounds that a higher SFM count correlates with a higher likelihood, when in reality, perhaps it is only the SFMs contained within the SFL itself which are important in this predictor. In that case, *time* and *fine* are no more likely to occur with <yC> on the grounds of stem-final <im> and <n> than are any other lexels with stem-final <m> or <n>.

Other lexels for which the likelihood of <yC> is predicted too low by Model 7 include *bigging* and *heir*. Tokens of these lexels never feature SFMs, so the fact that they are low outliers here indicates that they have an unusually low likelihood of <yC> than other non-minim-final tokens. A possible explanation for this is that *heirs* and *biggings* commonly occur with stem-medial <y-, as shown in (33).

(33) a. <ayris> *heirs* [text 74: 1408, CHA, BWK]

b. <byggy(n)nis> *biggings* [text 145: 1419, CHA, ROX]

9.3 Covered inflectional <i> and <y> in vpt and vpp {D} inflections

The left-hand pane in figure 9.12 shows the percentage of tokens per lexel for vpt and vpp {D} realised as <yC>. The right-hand pane shows the same variable on a per-text basis. The four boxplots shown in these two panes are very similar to one another, suggesting that there is unlikely to be a difference either between vpt and vpp or between individual text and individual lexel when it comes to likelihood of <yC>. In the following descriptive plots, therefore, there will be no distinction made between the two grammels.

FIGURE 9.12: Percentage of vpt and vpp {D} tokens realised as <yC>. The left-hand plot is based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plot is based on the percentage by text (so individual data points represent individual texts). Only texts and lexels with more than 10 tokens are included.

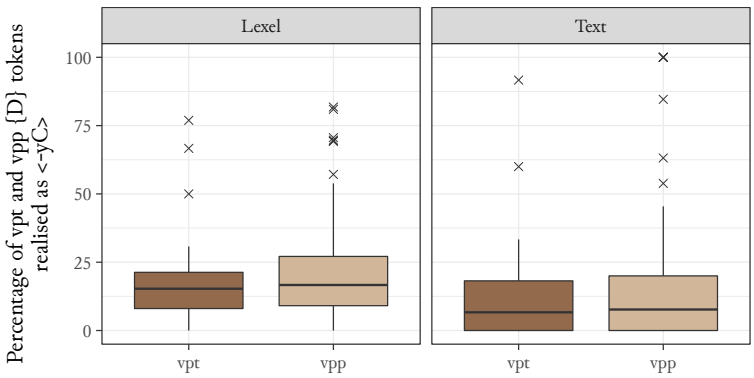


Figure 9.13 shows a negative correlation between date and percentage of vpt and vpp {D} tokens realised as <yC>. Whilst the paucity of tokens in the earliest years of the period covered by LAOS cause fairly wide confidence bands, these do not appear to render the negative correlation insignificant, though it is potentially a very weak correlation.

FIGURE 9.13: The percentage of vpt and vpp {D} tokens of each lexel in INFLAOS realised as <yC> as opposed to <iC> between 1380 and 1500. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Red: linear trend line; blue: smooth trend line.

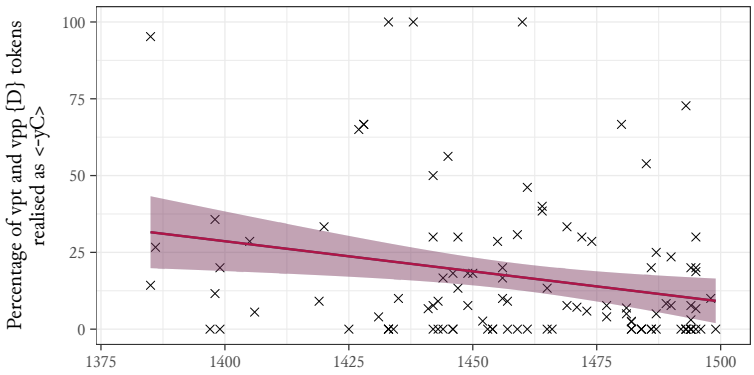


Figure 9.14 shows the percentage of vpt and vpp {D} tokens in INFLAOS texts of each type realised as <yC> as opposed to <iC>. The mean percentage values for burgh records (BUR), charters (CHA) and state documents (STA) are very similar - around 10%, whereas the mean percentage for cartularies (CAR) is higher, at approximately 30%, and for notarial protocol books is lower - close to 0%. The range of percentage values is widest for charters: the interquartile range is from 0% to 75%, and there are three texts which are outliers, with over 95% of tokens realised as <yC>. A similar pattern was shown in figure 9.5, which showed the percentage of npl {S} tokens realised as <yC>, whereby the widest range of percentage values was shown for charters. As noted in section 4.1.5, 'charter' as a type category is a kind of 'catch-all', encompassing a wider variety of manuscripts than the other, more specific categories. It may be that this masks some more specific type variation. This is something that is remedied to an extent by including the random effect of text in the generalised additive model formula as discussed in sections 9.2.1 and 9.3.1.

Figure 9.15 shows the correlation between number of SFMs and percentage of tokens with stem-final <yC> as opposed to <iC>. Each point on the graph represents an individual lexel. Because some lexels can be realised orthographically with different numbers of SFMs, some lexels are represented by multiple points, one for each SFM count value.

The positive correlation shown in figure 9.15 between SFM count and percentage of tokens realised as <yC> appears very similar to the positive correlation between the same variables for npl, shown in figure 9.8. That is, an increasingly high percentage of <yC> as SFM count increases from one to three. The fact that these graphs appear similar is evidence in favour of the hypothesis that OSc scribes' use of <y> as the CIV

FIGURE 9.14: The percentage of vpt and vpp {D} tokens in INFLAOS texts of each type realised as <yC> as opposed to <iC>. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown.

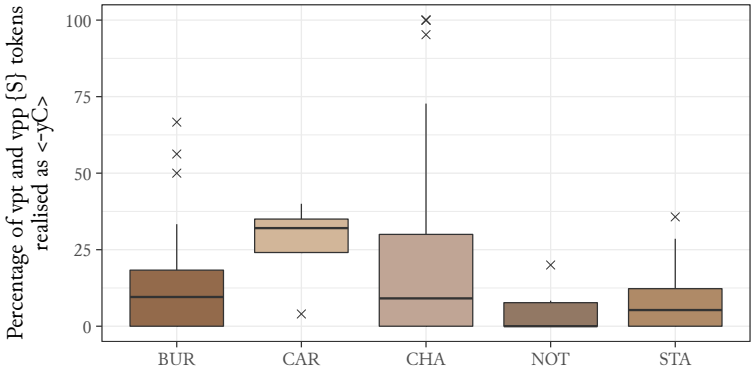
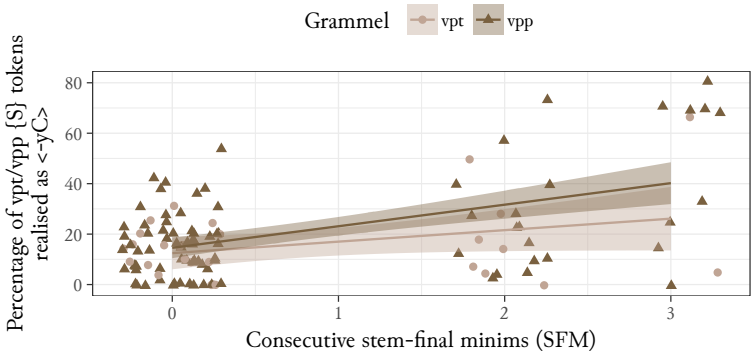


FIGURE 9.15: The correlation between number of stem-final minims (SFMs) and percentage of tokens with stem-final <yC> as opposed to <iC>. Each point represents an individual lexel. Only lexels exemplified by more than 10 tokens in INFLAOS are shown.



instead of <i> is, like their use of the <ſ> representation of {S}, a palaeographical phenomenon.

9.3.1 Modelling the likelihood of covered inflectional <y> in vpt and vpp {D}

The summary table of Model 8 shows the significant predictors remaining in the final GAM of the likelihood of vpt and vpp {D} being realised as <yC>. The contrast between the results of this model and Model 7 is striking. Model 7, predicting npl covered inflectional <y>, was reduced from a full model containing all potentially significant predictors to one which showed that the deviance within the INFLAOS npl data could be explained just as well using only the PV of SFM together with the random effects of text and lexel. In contrast, Model 8 retains both contextual fixed effect PVs (date, text type and location) and the lexical fixed effect PV of stem syllable count in addition to consecutive SFM.

Figure 9.16 is a heat map showing the relative log-likelihood of vpt and vpp {D} to be realised as <yC> depending on text location. The areas of high and low likelihood appear to map on to certain counties

Model 8: GAM of the likelihood of vpt and vpp {D} to be realised as <yC> as opposed to <iC>, predicted by: number of consecutive SFM; stem syllable count; text type; text date; and text location; and including random effects of individual text and lexel. **R^2 = 0.554; Deviance explained = 55.8%; N = 4,929**

Parametric coefficients:

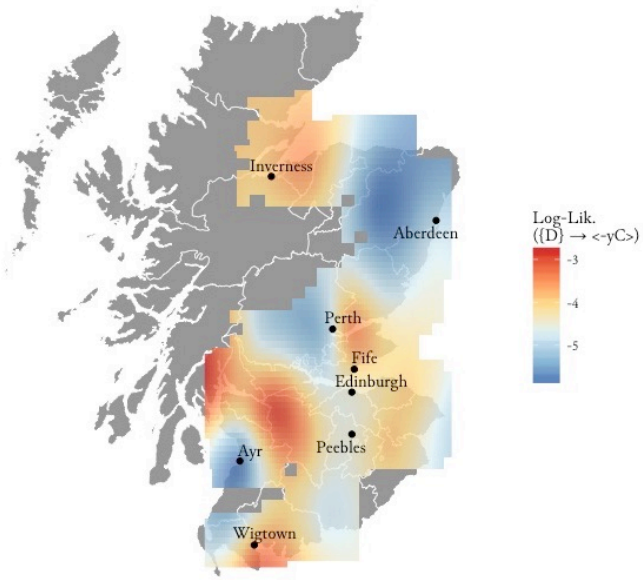
		Estimate	SE	z	
(Intercept)		-1.66	0.21	-7.92	***
Type	BUR	0.01	0.25	0.04	
	CAR	0.08	0.38	0.22	
	NOT	-3.96	0.68	-5.86	***
	STA	-2.26	0.46	-4.89	***
SFM	2	1.22	0.19	6.39	***
	3+	2.79	0.25	11.16	***
Syllables	Polysyllabic	-0.76	0.17	-4.35	***

Approximate significance of smooth terms:

	edf	Ref.df	χ^2	
Text	350.17	1007.00	928.58	***
Lexel	74.76	408.00	229.24	***
Date	5.62	6.17	22.44	**
Latitude, Longitude	18.83	20.97	57.68	***

with quite defined boundaries. For example, Aberdeen is shown to be an area with low likelihood of <yC>, whereas Inverness contrasts as a high-likelihood area. The south west is a high-likelihood area, with the exception of Ayr, which stands out as a particularly low-likelihood area.

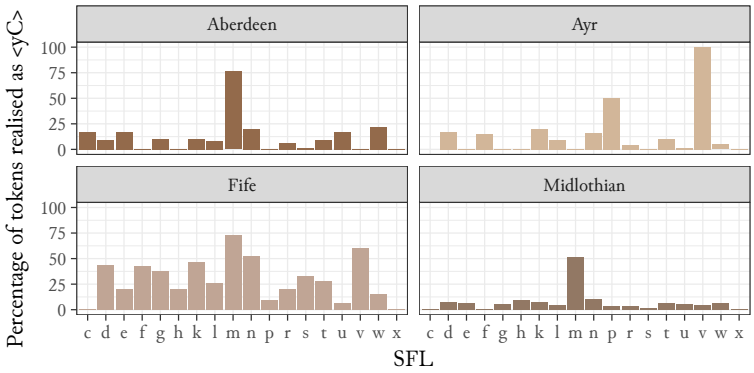
Figure 9.16: A heatmap showing the relative log-likelihood of vpt and vpp {D} being realised as <yC> over the geographical area represented by LAOS.



Given that the equivalent model for npl showed that the likelihood of <yC> could be successfully

predicted based on the single PV of stem-final minim count (controlling for individual text and lexel effects), it is surprising that the likelihood of <yC> in vpt and vpp inflections should be affected by a contextual PV such as location. Model 8 also controls for the random effect of individual text, but the correlation with specific geographic locations may point to stylistic variation common to scribes in particular areas. Figure 9.17 shows the percentage of vpt and vpp {D} tokens realised as <yC> for each SFL in (a) two counties which appear as areas of very low likelihood in figure 9.16 (Aberdeen and Ayr); (b) one county which appears as an area of medium likelihood (Midlothian); and (c) one county which appears as an area of high likelihood (Fife).

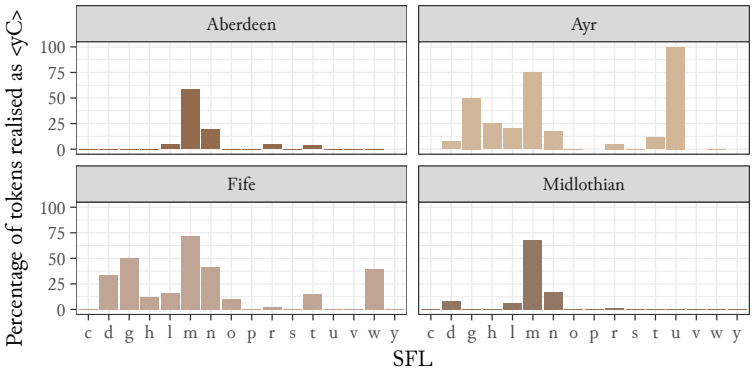
FIGURE 9.17: The percentage of vpt and vpp {D} tokens realised as <yC> for each stem-final *littera* (SFL) in Aberdeen, Ayr, Fife and Midlothian.



Aberdeen and Midlothian show the kind of distribution which might be expected if the likelihood of vpt and vpp <yC> patterned like that of npl <yC> . That is, the highest percentage of <yC> is attested following the SFL which are composed of the most minim strokes - <n> and <m>, and a low level of <yC> elsewhere. Ayr does not show this correlation with SFM, but this is due to there being only two <m>-final vpt or vpp tokens attested in Ayr texts (compared to 42 in ABD, 113 in FIF and 39 in MLO). Fife, the area of high likelihood in figure 9.17, has a high percentage of <yC> following consecutive stem-final minims, similar to Aberdeen and Midlothian, but displays a higher percentage of <yC> following most other SFL than the other three counties. This suggests that a scribal tendency existed in this area towards using <yC> rather than <iC> for reasons beyond its palaeographic utility.

Interestingly, the equivalent plot of the percentage of <yC> in npl {S} for these four counties, shown in figure 9.18, does not show any evidence of this trend, as suggested by the different conclusions of Models 7 (npl) and 8 (vpt and vpp).

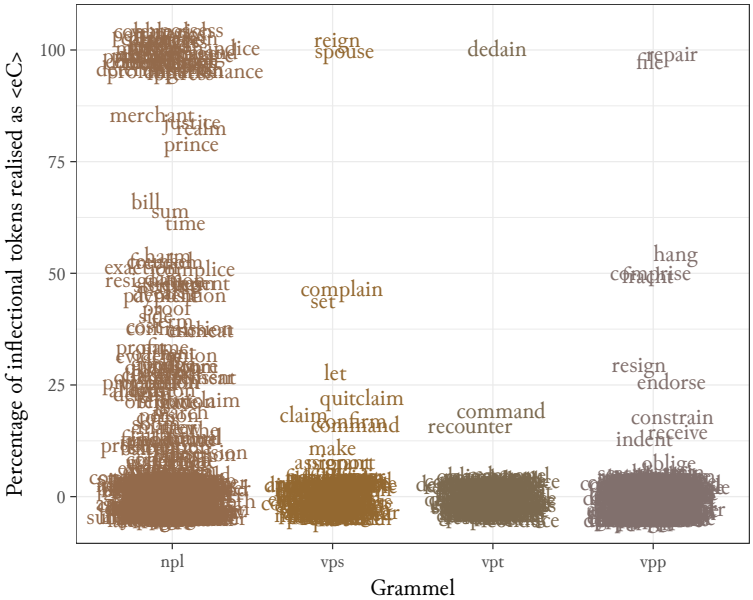
FIGURE 9.18: The percentage of npl {S} tokens realised as <yC> for each stem-final *littera* (SFL) in Aberdeen, Ayr, Fife and Midlothian.



9.4 Covered inflectional <e>

Figure 9.19 shows, for each grammel, all lexels in INFLAOS exemplified by 10 or more tokens with a <VC> inflectional form. The position of each lexel on the y-axis indicates the percentage of tokens of that lexel where the inflection is realised as <eC>. Whilst there are a few verb lexels which exhibit <eC> inflection, the vast majority of <eC> is found among npl tokens. The npl lexels in figure 9.19 form a large cluster between 0% and 50% <eC>, and another close to 100% <eC>. This suggests that whilst <eC> was used variably with some lexels, the lexels at clustered at the higher end of the scale form a subset which is particularly prone to <eC> realisation of {S}.

FIGURE 9.19: Realisation of npl, vps, vpt and vpp inflections as <eC> forms. Only lexels with >10 tokens included.



King (1997: 160) suggests that covered inflectional <e> occurs sporadically in OSc, mainly in the "earliest texts". She suggests that this is indicative of ME scribal influence on OSc scribal practice. In this section, I investigate this claim by considering the proportion of npl {S} tokens realised as <eC> in INFLAOS texts over time, as well as the proportion of INFLAOS texts using <eC> over time.

The extremely low attestation of CIV <e> in grammels other than npl means that it is not practicable to investigate its attestation in these grammels. Rather, the investigation into covered inflectional <e> presented here focuses solely on npl {S}. Section 9.4 is an investigation into the claim put forward by King (1997: 161) that occurrence of covered inflectional <e> is concentrated in the earliest OSc texts. I fit a statistical model to the INFLAOS npl data to determine whether text date is the defining factor in OSc scribes' use of covered inflectional <e>. The implication of King's statement that OSc covered inflectional <e> is confined to the earliest texts is this variant was due to ME influence on OSc scribal practice. To investigate whether this is the case, I select several linguistic features which have distinct forms characteristic of OSc texts of ME texts, and compare the use (or lack thereof) of OSc variants in a selection of early texts which have the highest percentage of npl covered <e> with the use of these variants in a selection of texts which use only <i> and <y> forms of npl {S}.

Whereas previous chapters have been focussed on identifying factors conditioning the realisation of particular inflectional variants in INFLAOS, the following investigation of covered inflectional <e> will also acknowledge the potential ambiguity of the inflection form itself. In section 4.1.3, I outlined this issue from a methodological point of view, showing that it is often unclear where the morphological boundary between stem and inflection should be placed in cases where:

(a) the noun stem is consonant-final; and

(b) the inflectional consonant is preceded by an <e>.

Section 9.4.1 presents an attempt to address this methodological difficulty by investigating not only the factors which predict <e> in npl tokens, but also those which predict final <e> in singular noun tokens. To do this, a subset of nouns are selected from LAOS which are attested in both singular and plural forms. The occurrence of 'inflectional' and final <e> is compared using contextual and lexical factors, as well as with regard to the potential etymological conditioning factors for stem-final <e>.

Figure 9.20 shows the percentage of fully-realised npl {S} inflections with covered inflectional <e> as opposed to covered inflectional <i> or <y>. The mean percentage for both tokens grouped by individual text and by individual lexel is very similar. On average, fewer than 10% of npl {S} inflections are realised as <eC>. A notable difference, however, is that many more individual texts appear as outliers at the higher end of the scale than individual lexels. This suggests that the use of covered inflectional <e> is conditioned by scribal choice, and therefore is more likely to show correlation with contextual rather than lexical PVs.

FIGURE 9.20: The percentage of npl {S} tokens realised as <eC>. The left-hand plot is based on the percentage by lexel (so individual data points represent individual lexels); and the right-hand plot is based on the percentage by text (so individual data points represent individual texts). Only texts and lexels with more than 10 tokens are included.

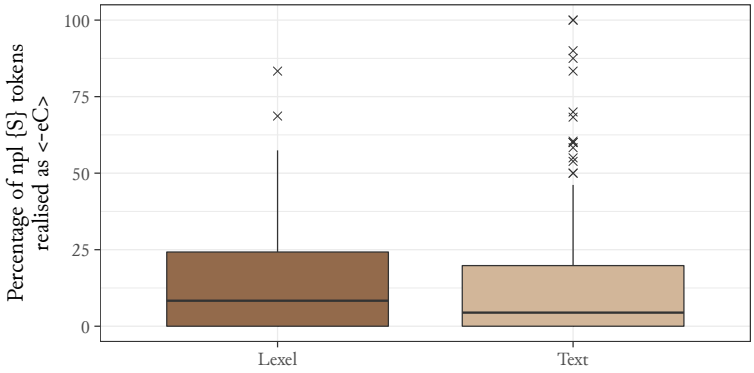
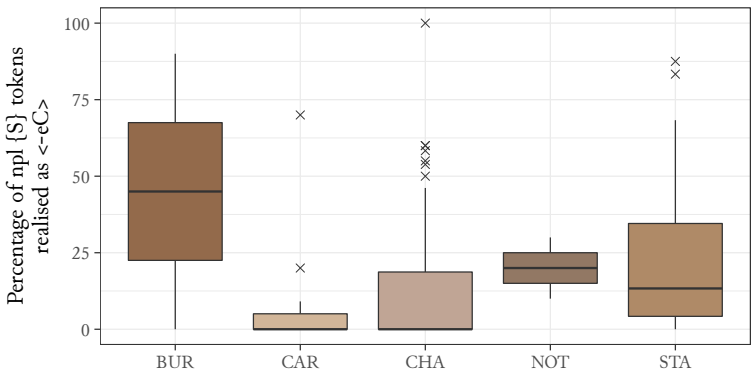


Figure 9.21 shows the percentage of npl {S} tokens in texts of each type realised as <eC> in INFLAOS. Burgh records (BUR) have the highest mean percentage of <eC> (approximately 45%), whereas cartularies (CAR) and charters (CHA) have the lowest (close to 0%). Burgh records and state documents (STA) show the most inter-text variation, as shown by their wider interquartile ranges. This may point to covariance between the contextual PVs of type, date and location. In particular, burgh records are extremely varied in terms of where and when they are attested. The fact that there is a lot of variation between individual burgh records with regard to <eC> suggests that these other factors are likely to have an effect. Section 9.4.1 presents a GAM including these PVs to ascertain whether these contextual PVs are significantly correlated with the presence of <eC> when considered together and in the presence of lexical PVs and random effects.

FIGURE 9.21: The percentage of npl {S} tokens in texts of each type realised as <eC> in INFLAOS.. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown.



9.4.1 Modelling covered inflectional and stem-final <e> in singular and plural nouns

Factors which may influence the occurrence of final <e> in singular nouns include:

(a) **Etymologically justified stem-final vowel**

Aitken and Macafee (2002: 70) found that in the *Scone Gloss* (LAOS text 160[1]), the occurrence of final <e> in singular nouns is not predictable based on etymology. Examples (34) and (35) are npl tokens of the lexel *side* from the *Scone Gloss*, derived from a vowel-final OE word, and *year*, derived from a consonant-final OE word. Both of these tokens are realised in the *Scone Gloss* with final <e>.

(34) <side> (*side* ← OE *sīde*)

(35) <iere> (*year* ← OE *gēr*)

(b) **Medial vowel length**

Minkova (2014: 189) describes the decline of the inflectional system from OE to ME, during which a weakening and neutralisation of vocalic inflections took place. The proliferation of orthographic <e> in inflectional syllables in the late OE and early ME periods points to a neutralisation to schwa, and the eventual loss of any phonetic value for such segments allowed the reanalysis of final <e> as a length diacritic.

(c) **Orthographic flourish**

Simpson (1973: MS12; n.3) refers to the tendency for OSc scribes to add a final <e> in environments where it is not justified either etymologically or as a length diacritic. In particular, he notes that a final <e> could be used by scribes to fill a gap at the end of a line to improve the overall appearance of a page of text by justifying the lines.

To investigate the occurrence of <e> in both singular-noun-final and npl {S} contexts, I use a dataset of singular nouns extracted from LAOS, and a subset of the plural nouns included in INFLAOS. These were chosen by lexel, with the criteria for inclusion being: (a) the lexel should have 30 or more tokens in each of the singular and plural forms; and (b) the lexel should be monosyllabic. The 30-token stipulation was to ensure that the analyses performed on the singular and plural datasets, and the conclusions drawn from them, would be comparable to one another, and the monosyllabicity stipulation to remove the necessity of ascertaining the OSc stress pattern of the lexel. Table 9.3 lists the 29 eligible lexels, and the corresponding number of tokens attested for each in the singular and in the plural.

These tokens were then categorised according to the PVs assigned in the creation of INFLAOS, and also assigned a value 'long' or 'short', according to the length of the medial vowel. These length judgements were based on Aitken's (2002) system of OSc vowels.

Model 9 attempts to predict the likelihood of stem-final <e> in singular tokens of the 29 noun lexels selected from LAOS. The resulting coefficient values suggest that contextual PVs are significant in predicting

[1]As noted in section 4.1.2, tokens from the *Scone Gloss* are not included in INFLAOS because it is dated 1360, 20 years earlier than the next earliest text.

TABLE 9.3: The 29 monosyllabic lexels with 30 or more tokens in both the singular and the plural in LAOS.

Medial V	Etymology	Lexel	Plural	Singular	Total
Long	Non-Germanic	march	54	34	88
		cause	43	123	166
		charge	32	50	82
		claim	38	101	139
		court	61	870	931
		faith	35	162	197
		heir	1,582	238	1,820
		judge	50	44	94
		part	49	1,288	1,337
		seal	110	521	631
		sum	67	470	537
		term	334	186	520
Long	Germanic	good	274	43	317
		lord	197	1,830	2,027
		mail	75	105	180
		name	39	163	202
		oath	32	133	165
		time	47	830	877
		year	796	1,137	1,933
Short	Non-Germanic	cost	94	33	127
		place	43	265	308
		rent	76	139	215
Short	Germanic	hand	354	236	590
		knight	31	144	175
		land	1,342	1,372	2,714
		right	88	220	308
		scathe	71	66	137
		son	32	412	444
		thing	164	94	258
Total			**6,210**	**11,309**	**17,519**

the likelihood of stem-final <e>. Specifically, burgh records, cartularies and state documents more likely to contain stem-final <e> than charters or notarial protocol books. Text location is also a significant predictor of final <e>.

Model 10 predicts the likelihood of npl {S} to be realised as <eC> in the subset of INFLAOS tokens shown in table 9.3. It includes the same significant PVs as Model 9.

9.4.1.1 Text type

Figure 9.22 shows the log-likelihood in each text type of npl covered inflectional <e> estimated by Model 10 (left-hand plot) and of final <e> in singular nouns estimated by Model 9 (right-hand plot). The relative likelihoods of covered inflectional <e> and of singular final <e> are similar for cartularies (CAR), charters (CHA) and notarial protocol books (NOT), with cartularies least likely, and notarial protocol books most likely out of these three types to exhibit both types of <e>. Burgh records (BUR), are the least likely texts

Model 9: GAM modelling the likelihood of stem-final <e> in a subset of singular nouns. **R^2 = 0.489; Deviance explained = 46.3%; n = 10,204**

Parametric coefficients:

		Estimate	Std. Error	z-value	
(Intercept)		0.43	0.64	0.67	
Type	BUR	-0.93	0.15	-6.27	***
	CAR	-0.94	0.22	-4.33	***
	NOT	0.36	0.22	1.65	.
	STA	1.07	0.24	4.55	***
Medial V	Short	-2.82	0.98	-2.88	**
Etymology	Gmc	-0.39	0.92	-0.43	

Approximate significance of smooth terms:

		edf	Ref.df	Chi.sq	
Text		471.04	1014.00	2566.36	***
Lexel		24.92	26.00	2433.73	***
Date * Etymology	Non-Gmc	1.03	1.03	0.01	
	Gmc	6.44	7.31	47.19	***
Date * Medial V	Long	3.64	4.26	18.66	**
	Short	4.47	5.34	24.16	***
Latitude, Longitude		8.31	9.33	58.45	***

Model 10: GAM modelling the likelihood of stem-final <e> in a subset of INFLAOS npl {S} tokens. **R^2 = 0.78; Deviance explained = 79.4%; n = 5,575**

Parametric coefficients:

		Estimate	Std. Error	z value	
(Intercept)		-3.51	0.75	-4.69	***
Type	BUR	0.28	0.43	0.65	
	CAR	-1.47	0.59	-2.51	*
	NOT	1.64	0.56	2.93	**
	STA	1.42	0.63	2.26	*
Medial V	Short	-1.94	1.17	-1.67	.
Etymology	Gmc	-1.38	1.08	-1.28	

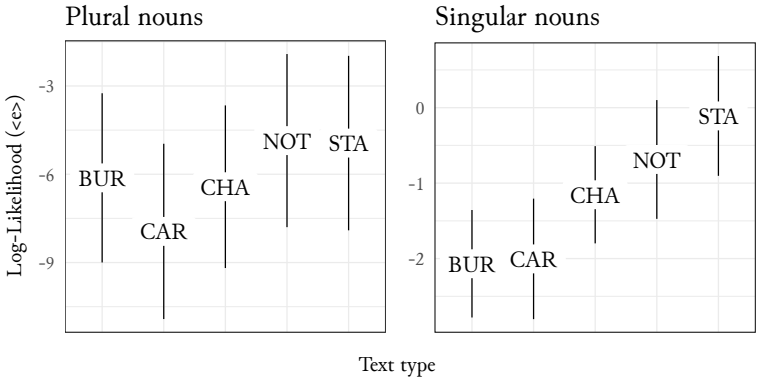
Approximate significance of smooth terms:

		edf	Ref.df	Chi.sq	
Text		131.81	871.00	355.43	***
Lexel		20.76	26.00	604.66	***
Date * Etymology	Non-Gmc	3.20	3.67	13.98	**
	Gmc	5.68	6.54	20.02	**
Date * Medial V	Long	0.00	0.00	0.00	
	Short	1.00	1.00	8.67	**
Latitude, Longitude		9.15	10.99	62.57	***

to contain singular final <e>, but this correlation is not mirrored for covered inflectional <e>, which is least likely to occur in cartularies, whilst its likelihood in burgh records is similar to that in charters, but not quite as high as in notarial protocol books or state documents. State documents are the most likely text type to contain singular final <e>, but do not stand out so much compared to other types when it comes to

likelihood of covered inflectional <e>.

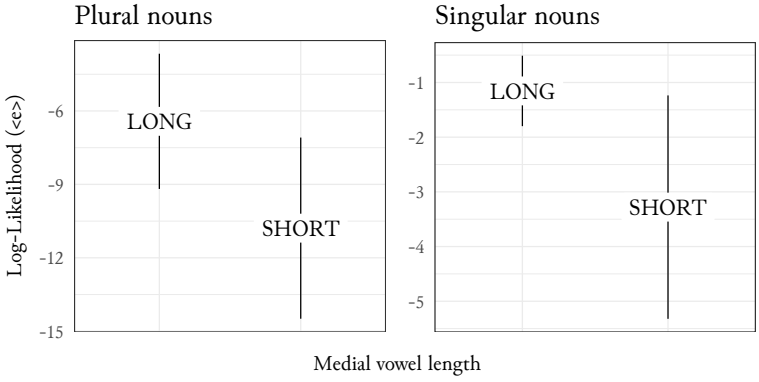
FIGURE 9.22: The log-likelihood of <e> in different text types estimated by Model 10 (left-hand plot - covered inflectional <e> in plural nouns) and Model 9 (right-hand plot - final <e> in singular nouns).



9.4.1.2 Medial vowel length

Figure 9.23 shows the log-likelihood of npl covered inflectional <e> and of final <e> in singular nouns according to the medial vowel length of the monosyllabic stem. In singular nouns, final <e> is more likely following a long medial vowel. This suggests that <e> functioned as a length diacritic in the LAOS texts. The confidence interval for the category of long medial vowels is very narrow compared to the interval for short medial vowels, suggesting that, whilst stem-final <e> was overtly likely to be found following long medial vowels, it was also found to variable degree following short vowels. In other words, if we were to predict whether a given token had a final <e> based only on its medial vowel length, knowing that the token had a long medial vowel would be a good indicator that the token had a final <e>, whereas knowing that the token had a short medial vowel would not really help to guess at the presence or absence of final <e>.

FIGURE 9.23: The log-likelihood of <e> depending on medial vowel length estimated by Model 10 (left-hand plot - covered inflectional <e> in plural nouns) and Model 9 (right-hand plot - final <e> in singular nouns).



For npl covered inflectional <e>, the same trend is apparent - long medial vowels are more likely to be followed by <eC> than short medial vowels. The confidence intervals shown here are much more equal than those in the singular noun plot, suggesting variability in <e> occurrence following both long and short medial vowels, notwithstanding the overall higher likelihood of <eC> following long vowels. It may be that this higher variability in the presence of the same overall correlation indicates that the correlation of medial vowel length with <eC> is an indirect product of this correlation in singular nouns. That is, the likelihood of stem-final <e> for a particular lexel may be a contributing factor to its likelihood of covered inflectional <e>, without the actual medial vowel length having an effect on the CIV.

Model 10 has an R^2 value of 0.78, indicating that the combination of PVs it represents is able to account for almost 80% of the overall deviance (variation) in the subset of plural nouns. In other words, the model is a very good fit to the data. By contrast, Model 9, though it looks similar to Model 10 in terms of the PVs which are found to be significant, explains less than 50% of the overall deviance in the subset of singular nouns (R^2 = 0.48). This means that Model 9 is not a good fit to this subset of data overall. This could indicate that there are other factors which correlate with the occurrence of final <e> in singular nouns which have not been included in the model, or it could simply be that final <e> occurs more randomly than covered inflectional <e> in OSc.

This latter option is certainly a possibility given the tendency described by Simpson (1973: MS12; n.3) of OSc scribes to add <e> to the end of words where it did not encode any meaning. Having said that, the only explicit reason for this practice Simpson gives is to improve the appearance of a page of text by justifying the lines. Intuitively, it seems plausible that the presence or absence of an optional SFL such as <e> would be utilised in this manner. However, as shown by table 9.4, this is not evidenced by the subset of singular nouns investigated here. Of the 10,204 singular noun tokens in the subset, 615 occur immediately before a line break. Of those, 151 (24.6%) end in <e>. Of the remaining 9,589 tokens which occur line-initially or line-medially, 2,864 (29.9%) end in <e>. These figures suggest that final <e> is actually found less often before a line-break than elsewhere. This does not mean that scribes did not utilise <e> as a gap-filler in the way Simpson suggests, but it suggests that this is not a crucial explanatory factor in the occurrence of final <e>.

Table 9.4: The relationship between a line-final token position and the occurrence of stem-final <e>.

SFL	Pre-line break	Elsewhere	Total
<e>	151 (24.6%)	2,864 (29.9%)	3,015 (29.5%)
Other	464 (75.4%)	6,725 (70.1%)	7,189 (70.5%)
Total	615 (100%)	9,589 (100%)	10,204 (100%)

9.4.1.3 Date, etymology and medial vowel length

Figure 9.24 compares the log-likelihood of <e> over time for Germanic and Non-Germanic tokens estimated by Model 10 (left-hand plot: covered inflectional <e> in plural nouns) and Model 9 (right-hand plot: final <e> in singular nouns). The likelihood of covered inflectional <e> in non-Germanic tokens does not appear to vary much between 1380 and 1500, apart from a slight increase in likelihood in the final quarter of the fifteenth century. Final <e> in non-Germanic singular nouns appears to experience a dip towards the middle of the period, but the extremely wide confidence intervals mean that this is not a reliable trend to draw conclusions from.

FIGURE 9.24: The log-likelihood of <e> over time for Germanic and Non-Germanic tokens estimated by Model 10 (left-hand plot - covered inflectional <e> in plural nouns) and Model 9 (right-hand plot - final <e> in singular nouns).

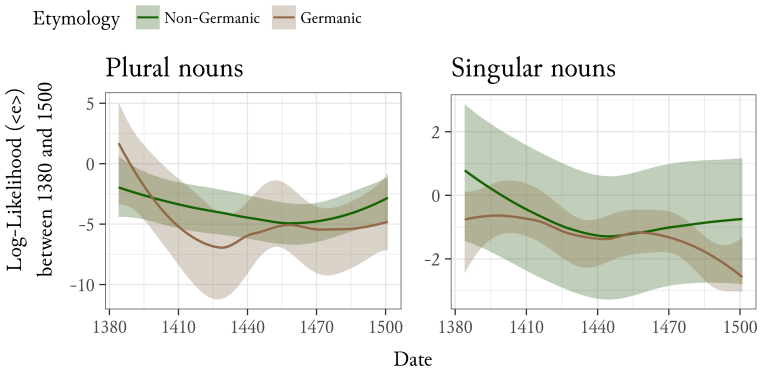
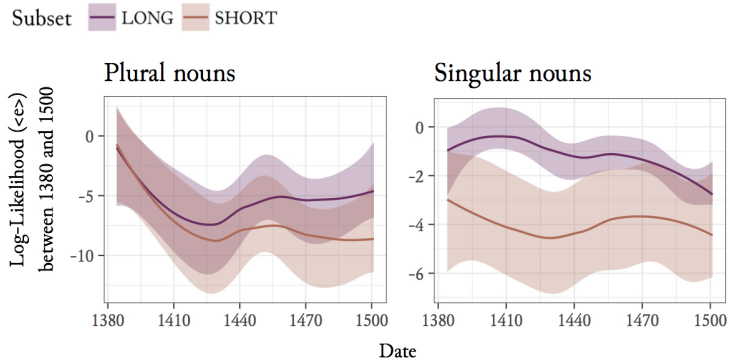


Figure 9.24 indicates that covered inflectional <e> in Germanic tokens is more variable over time than in non-Germanic tokens. The likelihood of an npl CIV being realised as <e> is highest at the beginning of the period 1380-1500, and declines sharply between 1380 and 1430. However, after 1430, the likelihood of <e> as the CIV appears to increase slightly and remains fairly static until 1500. By contrast, final <e> in singular Germanic nouns shows a fairly steady decrease between 1380 and 1500, with a steeper drop in likelihood observed at the very end of the period, between 1470 and 1500.

Figure 9.25 compares the log-likelihood of <e> over time for tokens with long and short medial vowels estimated by Model 10 (left-hand plot - covered inflectional <e> in plural nouns) and Model 9 (right-hand plot - final <e> in singular nouns). Whereas the likelihood of final <e> in singular nouns appears extremely conditioned by the length of the stem-medial vowel in the first half of the period, the likelihood of covered <e> in npl tokens shows no difference with regard to this variable. This changes in the second half of the fifteenth century, however, with long medial vowels appearing more likely to be followed by covered <e>. In the singular, the opposite trend can be observed, with the relative likelihood of final <e> following a long medial vowel and following a short medial vowel begin to converge.

FIGURE 9.25: The log-likelihood of <e> over time for tokens with long and short medial vowels estimated by Model 10 (left-hand plot - covered inflectional <e> in plural nouns) and Model 9 (right-hand plot - final <e> in singular nouns).



9.4.2 The persistence of <eC> over time

Figure 9.26 shows the percentage of npl {S} tokens realised as <eC> per text. A linear regression line (red) plotted through the data points suggests a gradual overall decrease in the percentage of <eC> tokens, but a smooth line gives a different impression. King's (1997) suggestion that the earliest texts use <eC> is corroborated. All but one texts dated before 1400 have <eC> as the realisation of at least 20% of their npl {S} tokens. Having said that, the use of <eC> in the later part of the period up to 1500 is not uniformly low. As shown by the smooth trend line, the percentage of tokens with <eC> per text troughs between 1425-1450, a period which is represented by few texts (see section 4.1.5), before rising again. The average percentage of <eC> in texts is undeniably far higher before 1400 than at any other time, but the sheer number of texts in the second half of the period compared to the first is far higher.

FIGURE 9.26: The percentage of npl {S} tokens of each lexel in INFLAOS realised as <eC> between 1380 and 1500. Each point represents an individual text. Only texts exemplified by more than 10 tokens in INFLAOS are shown. Red: linear trend line; blue: smooth trend line.

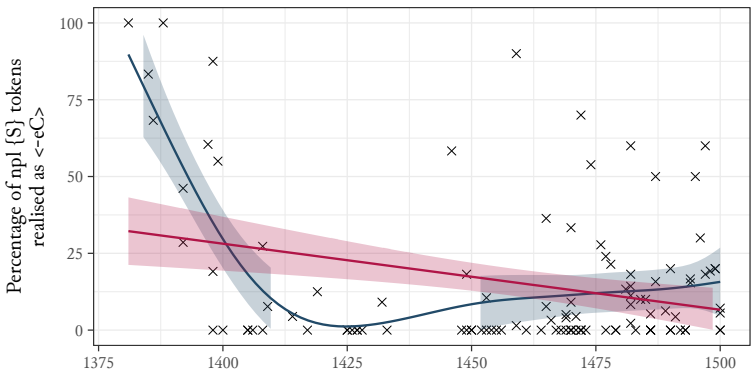
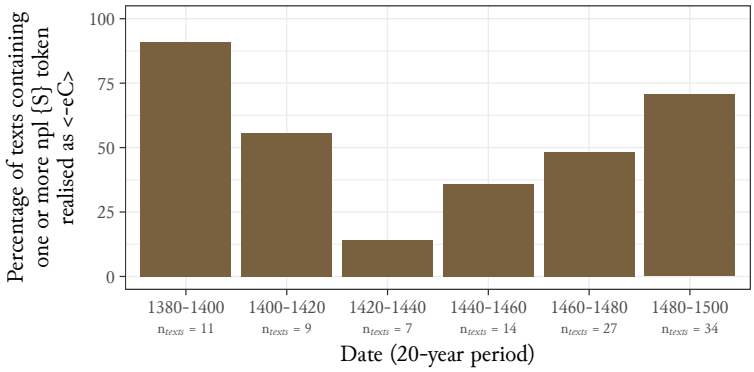


Figure 9.27 represents the same texts as those shown as individual data points in figure 9.26 (texts with 10 or more npl {S} tokens), this time grouped by date into 20-year periods. The bar plot shows, for each

20-year period, the percentage of texts containing at least one token of npl {S} realised as <eC>.

FIGURE 9.27: The percentage of texts containing at least one token of npl {S} realised as <eC> for each 20-year period in LAOS



As suggested by King (1997: 160) and by the INFLAOS evidence in figure 9.26, figure 9.27 shows that the period with the highest percentage of <eC>-containing texts is the earliest: 1380-1400. However, figure 9.27 also corroborates the apparent increase in <eC> usage in the second half of the period. 24 out of 34 (71%) of texts dated between 1480 and 1500 contain one or more npl {S} <eC> token, compared to 10 out of 11 (91%) of texts dated between 1380 and 1400. These results suggest that, contrary to King's assertion, covered inflectional <e> in OSc was not completely confined to the earliest texts.

This conclusion does not, however, disprove the idea that <eC> in the earliest texts was due to ME influence. Whether this was the case is difficult to test, but it is possible to examine the two periods of highest <eC> usage, 1380-1400 and 1480-1500, to see if the distribution is different, and therefore potentially motivated by different factors.

FIGURE 9.28: Wordclouds showing the lexels with the highest percentage of npl {S} tokens realised as <eC> in the periods 1380-1400 and 1480-1500.

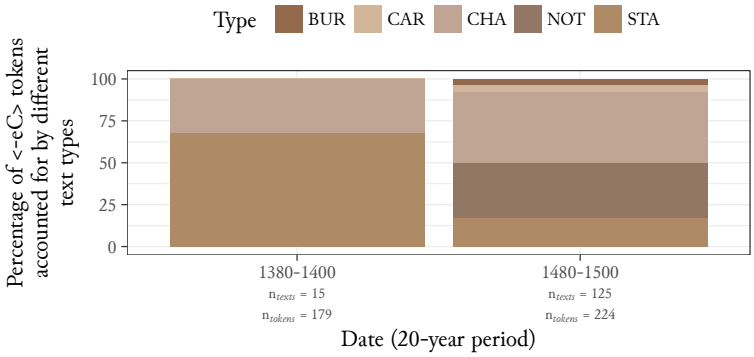
(A) 1380-1400

(B) 1480-1500

Figure ?? shows the percentage of total npl {S} <eC> tokens in INFLAOS attested for each text type

in the periods 1380-1400 and 1480-1500. In the earlier period, 68% of all <eC> tokens occur in state documents (STA), with the rest appearing in charters (CHA). In the later period, state documents contain only 17% of all <eC> tokens, with charters (42%) and notarial protocol books (NOT; 33%) accounting for the majority.

FIGURE 9.29: The percentage of total npl {S} <eC> tokens in INFLAOS attested in each text type for the periods 1380-1400 and 1480-1500.



It is clear that there is a difference in the distribution of <eC> tokens across text types between 1380-1400 and 1480-1500. Another noticeable difference shown by figure 9.29 is between the text and token n-values for each period. Between 1380 and 1400, there are 179 <eC> tokens attested in 15 texts. Between 1480 and 1500, a similar number of tokens, 224, is attested within 125 texts. This gives an average of 11.9 <eC> tokens per text between 1380 and 1400, and 1.8 per text between 1480 and 1500. These figures suggest that OSc scribes' use of <eC> npl {S} forms at the beginning of the fifteenth century compared to the end of the fifteenth century does not follow the same pattern.

Chapter 10

Discussion

10.1 {S} realised as zero

The main conclusion suggested by the initial descriptive analysis of the realisation of npl {S} tokens as zero in the INFLAOS data (section 6.3.1) was that, as suggested by Kopaczyk (as Bugaj (2002)), the occurrence of zero is predicted by the *litteral* and phonetic similarity between stem-final segments and fully-realised forms of {S}. This effect is observed most often in the realisation of the extremely common lexel *witness*, but also occurs in other *-es(s)*-final lexels such as *burgess* and *interest*, and *-is(s)*-final lexels such as *devise* and *premiss*. A GAM of the likelihood of npl {S} to be realised as zero (Model 1; page 103) supported this finding by attributing a significant amount of the deviance in the INFLAOS npl data to the variation between individual lexels, whilst simultaneously estimating stem-final <s> alone to be insignificant as a predictor of zero. Having said that, SFL did not emerge as an entirely insignificant predictor of zero according to Model 1. The presence of stem-final <ß> was shown to be a significant predictor of a zero-realisation of npl {S}. Based on the findings that zero is significantly more likely following stem-final <ß> but not following single stem-final <s>, the possibility that <ß> was itself acting as an inflectional form was considered. The hypothesis tested was that an npl form ending in <ß> such as *horses* in (36) might be a plural form contrasting with a singular form ending in a single <s>, such as *horse* in (37). This hypothesis was shown to be false by comparing the occurrence of stem-final <ß> in a subset of plural nouns assumed to have a zero inflection and occurrence of stem-final <ß> the corresponding singular nouns, such as *horse* in (38). The occurrence of stem-final <ß> was found to be very similar in both singular and plural realisations of the subset of lexels, suggesting that the increased likelihood of zero npl inflection following stem-final <ß> is due to a scribal aversion to following <ß> with a realised form of {S}.

(36) <Sexten oxyne four ky thre kalfꝑ twa horß>

'sixteen oxen, four cows, three calves, two horses' [text 704: 1486, CHA, FIF]

(37) <a hors the price of iiij-li$\bar{b}$>
'a horse, the price of four *Libra* [pounds]' [text 1675: 1447, BUR, ABD]

(38) <w^{t} fredomē to hald anȩ horß>
'with freedom to hold one horse' [text 265: 1498, CHA, ROX]

As well as the apparent correlations of stem-final {S}-like segments and stem-final <ß> with zero, the descriptive analysis indicated that text type correlated with the occurrence of zero, with notarial protocol books appearing more likely to contain zero-inflected npl {S} than other types. This correlation was supported by Model 1, which showed a significantly higher likelihood of zero in notarial protocol books compared to other types of text. Having said that, a qualitative analysis of the notarial protocol book texts contained in INFLAOS revealed that this increased tendency toward zero inflection can be traced not only notarial protocol books as a type, but further, to one specific notary, Sir Thomas Crawfurd. Even more, whilst a generally higher likelihood of zero in texts transcribed from Crawfurd's protocol book might suggest an element of scribal preference or idiosyncrasy on his part, the vast majority of npl zero tokens used by Crawfurd are instances of the lexel *pertinent*, a word which Crawfurd uses very often - far more often than either of the other two notaries whose protocol books appear in INFLAOS, James Young and Peter Marche. Crawfurd's high usage of zero is therefore a consequence of his high usage of the word *pertinents*, which may itself indicate a difference in the subject matter of the legal transactions he notarised, or alternatively simply a difference in his preferred phrasing.

Not only the type, but also the date and location of a text were shown to be significant predictors of the realisation of npl {S} as zero, though the correlations shown were not strong enough to draw firm conclusions like those regarding SFL and text type. These contextual predictors were, however, not significant in predicting the same phenomenon for vps {S}. This suggests that the elements of spatio-temporal variation which condition the zero-realisation of npl {S} are absent for vps {S}. An explanation for this is suggested by the significance of Northern Subject Rule (NSR) environment as a predictor of vps zero. Model 2 (page 114) showed that zero npl inflection of vps tokens is significantly more likely if the token occurs in an 'NSR environment'. In Model 2, a vps token was considered to occur in an NSR environment if it fulfilled both of the following criteria:

(a) the subject of the verb is a personal pronoun; and

(b) the subject of the verb is adjacent to the vps token.

These criteria are generally referred to as the 'type-of-subject' and 'proximity-of-subject' constraints respectively (Rodriguez Ledesma 2013). The reason for encoding this variable as a binary distinction was

shown in section 6.3.3, in which a preliminary analysis was made of the effect of these two constraints by plotting the percentage of vps tokens with a zero inflection in the following four environments (figure 6.14):

(a) the subject of the verb IS NOT a personal pronoun and IS NOT adjacent to the verb (i.e. neither the type-of-subject nor the proximity-of-subject constraint is fulfilled);

(b) the subject of the verb IS a personal pronoun but IS NOT adjacent to the verb (i.e. only the type-of-subject constraint is fulfilled);

(c) the subject of the verb IS NOT a personal pronoun but IS adjacent to the verb (i.e. only the proximity-of-subject constraint is fulfilled); and

(d) the subject of the verb IS a personal pronoun and IS adjacent to the verb (i.e. both the type-of-subject and the proximity-of-subject constraints are fulfilled);

This analysis showed that the vast majority of vps tokens with a zero inflection occur in an environment where both the both the type-of-subject and the proximity-of-subject constraints are fulfilled. Environments where only one of the constraints is fulfilled showed no meaningful difference in the occurrence of zero inflections compared to environments in which neither constraint was fulfilled. This seems at odds with the analysis of the NSR in Northern Middle English (NME) by Haas and Van Kemenade (2015) who finds that both constraints individually have a significant effect on vps inflectional forms. However, Haas and Van Kemenade (2015: 55) notes that in the more northerly "core" area of NSR operation, the interaction between the two constraints is more significant than in "peripheral" areas further towards the Midlands. That is, where the operation of the NSR is stronger, the combined effect of subject type and adjacency is stronger.

The pattern shown by the LAOS vps tokens suggests that in OSc, the operation of the NSR is conditioned only by the interaction between these two constraints, and that a non-adjacent pronominal subject, or an adjacent non-pronominal subject are not in themselves sufficient to trigger it. The majority of previous scholarship on the NSR has focussed on NME, but a study by Rodriguez Ledesma (2013) investigating the operation of the NSR in OSc also using the LAOS data finds a similar pattern in lexical vps tokens: 99.5% of tokens in environments where both conditions are met have a zero inflection, compared to 0.44% in environments where the subject is a non-adjacent pronoun. Rodriguez Ledesma's study focusses only on first-person vps forms, and there is only one occurrence in LAOS of a vps token with an adjacent, non-pronominal subject. However, these results clearly point to the same conclusion as my analysis of all relevent vps tokens in LAOS, that the NSR in OSc operated in environments where both conditions were met. Rodriguez Ledesma (2013: 169) suggests, based on a comparison of vps forms in *A Linguistic Atlas of Early Middle English* (LAEME), that "Scots was more advanced than Northern English in the implementation of the NSR". She further draws a parallel between the NSR and another phenomenon which is

attested in OSc earlier than in NME, the use of <i> as a length diacritic, to demonstrate that features being more "advanced" in OSc is observed elsewhere than the NSR.

The fact that the NSR is less variable in OSc than in NME suggests that it was more of an entrenched part of the verbal paradigm of OSc. The analyses of Haas and Van Kemenade (2015) and Rodriguez Ledesma (2013) taken together with the current study suggest significantly less variability in the operation of the NSR in OSc compared to NME. This regularity of NSR operation in OSc may explain why, as noted above, the occurrence of zero inflection in vps tokens does not show the same spatio-temporal variation as zero inflection in npl tokens. That is, a 'regularised' part of the grammar, such as the NSR in OSc has been shown to be, would be expected to be less susceptible to temporal and spatial variation. Whereas npl zero, as a variant inflection not subject to grammatical constraints, might well be expected to vary according to the preferences or conventions of a scribe or group of scribes, the status of vps zero, as a grammatically-conditioned variant, would be less susceptible to scribal preference or idiosyncracy.

Interestingly, despite the fact that vps zero is predicted with a significant degree of accuracy by whether the context in which it occurs is subject to the NSR, the presence of stem-final <ß> is also significant in predicting vps zero, as it was for npl zero. This suggests that, even though the zero realisation of vps {S} was a systematic part of the morpho-syntax of OSc, the scribal aversion to following stem-final <ß> with an orthographic form of {S} is still evident.

Neither npl nor vps {S} showed a significant amount of variation between individual texts, suggesting that the use of zero was not significantly influenced by scribal idiosyncrasy. This shows that the {S}-repelling effect of stem-final <ß> was a general scribal characteristic, rather than the stylistic choice of a few individuals. The variation between individual lexels, however, was significant for both npl and vps zero, suggesting that certain words were more likely to take a zero-inflection than others. This was true even in the presence of the main effect of SFL and, in the case of vps, the main effect of NSR. When this variation between individual lexels was investigated, however, it emerged that much of this variation could be traced to irregular forms. In particular, the third-person singular vps form of the lexel *will* is attested frequently with a zero {S} inflection in contexts where its subject is not an adjacent personal pronoun and therefore would not trigger the operation of the NSR. Examining the full lexical context of these *will* forms revealed that a large majority of them occurred as part of the set legal phrase *as the law will* (or slight variations thereof). This could be interpreted as a fossilised subjunctive form or, alternatively, might indicate a syntactic analysis of the noun phrase *the law* which required plural agreement as a mark of authority, as with the 'royal we' pronoun usage by monarchs and other authority figures.

10.2 Abbreviated {S}

Chapter 7 began with a descriptive comparison of the percentage of npl and vps {S} tokens realised as the abbreviation symbol <ſ> for different lexels and for different texts. This comparison indicated a large amount of variation between different lexels, spread across the whole of the percentage scale from 0% to 100%, with a mean of approximately 60% suggesting a higher number of lexels with a large proportion of tokens realised as <ſ> than lexels with a low proportion of <ſ> tokens. The distribution of percentage <ſ> in different texts showed less dramatic variation, with the mean percentage of npl tokens abbreviated per text matching the mean percentage across lexels. In other words, the variation in <ſ> is between different lexels, and the variation between texts is simply a consequence of the fact that different texts contain different numbers of particular lexels. If a lexel is highly likely to be inflected with <ſ>, and a text contains many tokens of that lexel, then it follows that the text will show a high level of <ſ>. However, the mean percentage of <ſ> in vps lexels is not matched as closely by the percentage of <ſ> in individual texts, as was the case for the equivalent npl figures. Instead, the overall percentage of vps <ſ> per text is lower, around 40%. This suggests that the realisation of vps {S} as <ſ> may be influenced to a degree by scribal choice. A possible explanation for this might be that, because the OSc vps {S} paradigm was subject to a systematic alternation between zero and fully-realised {S} in particular syntactic contexts due to the operation of the NSR, scribes could have had a more 'fixed' notion of the correct realisation of {S} in specific contexts. There was no such alternation operating within the npl paradigm, so scribes would not have a preconceived awareness of systematic variation in npl forms, resulting in a less constrained use of <ſ>.

Despite this apparent difference in scribal variation for npl and vps zero, there is no linguistic factor which would suggest that abbreviation should be more likely in npl as opposed to vps {S}. The model which is fitted to the data (Model 3; page 125) therefore conflates the data for both grammels, but retains the distinction between them as a PV. Another reason for this decision was that the INFLAOS data contains a much smaller number of vps tokens than npl tokens once those tokens realised with a zero inflection are removed (see). The difference in the likelihood of abbreviation was also reflected in the results of Model 3, which showed that, when all other significant predictors are taken into account, vps tokens are less likely to be inflected with <ſ> than npl tokens.

The predictor which shows the most dramatic correlation with the realisation of {S} as <ſ> is SFL. The SFL attested in the npl and vps tokens included in INFLAOS can be clearly separated into two distinct groups, those which are very often followed by <ſ> and those which are very rarely followed by <ſ>. The reason behind this grouping is immediately observable based on an examination of the *figural* representation of each *littera*. *Littera* which culminate in a horizontally-extending stroke are likely to be followed by <ſ>, and those which culminate in a vertical down-stroke are unlikely to be followed by <ſ>. The only *littera* which does not seem to fit precisely within one or other of these categories is <l>, for which two explanatory

factors are proposed. Firstly, that there are palaeographic variants of <l> such that it is possible to find <l> culminating in: (a) a horizontal stroke (when it is realised as a 'looped' form similar to the typical form of the ascender in <d>); or (b) a vertical stroke (when it is realised as a form resembling an elongated minim stroke).

As well as this correlation of {S} abbreviation with SFL, Model 3 also indicates an increase in the likelihood of <ſ> over time in the first half of the period covered by LAOS (1380 to around 1440), followed by a plateau in the likelihood of <ſ> in the remaining period to 1500. The plateau is explained by the correlation of <ſ> with SFL- the use of abbreviation can only reach a certain level because its usage is constrained by SFL. 100% abbreviation is impossible because it can occur only in particular palaeographic contexts. The increase in <ſ> over the first half of the period may indicate that, as legal texts were increasingly often written in the vernacular rather than in Latin, scribes developed and increased their use of time-saving palaeographic strategies.

Another factor shown by Model 3 to correlate with the use of <ſ> was text type. Specifically, burgh records were most likely to contain abbreviated forms of {S}, and state documents lest likely. This distinction seems intuitively reasonable given the formality and purpose of these types of text. Burgh records were used to keep note of the proceedings taking place in a burgh court on a day-to-day basis. Whilst they were intended as a record which could be referred back to, there is a clear difference between the status of, for example, a burgh record book used to record minor court proceedings such as the theft of hose described in the excerpt from the *Newburgh Court Book* in (39); and a state document recording the intention to make peace with England, as in the excerpt in (40). The purpose of a burgh court book was to record the decisions made and actions taken during each court session - a necessary record of who had been fined, imprisoned or acquitted of which crime on a particular day. It might be important to refer back to the record in (39) should John Malcomson dispute the ownership of the hose in question at a later date, but otherwise the record is of little consequence. On the other hand, the subject matter of (40), is an important development in the diplomatic relations between Scotland and England. A scribe recording this kind of announcement would be conscious that he was recording for posterity a document which had great significance. The difference between the writing of the hypothetical high court scribe and burgh court scribe in these scenarios is both the formality and the perceived longevity of the documents they are producing. It may be, then, that the burgh court scribe made more use of the time-saving method of abbreviation in writing what he perceived as a necessary, everyday document than the high court scribe writing a formal document to preserve an important decision.

(39) <ye-qu^{k} day John̄-malcomson in am(ercemen)t for he wrangwysly hyld fra John dauid-malcomson a p(er) of hoyß>
'The-which day, John Malcomson, in amerciament for he wrongwisely held from John David Mal-

comson a pair of hose'

On the same day, John Malcomson is fined for unlawfully withholding a pair of hose from John David Malcomson

(40) <It Is ordanit avisit and Concludit be oure souerane Lord and his thre estaitis being assemblit in this p(rese)nt parliame(n)t that pece be takin with Ingland>

it is ordained, advised and concluded by our sovereign lord and his three estates being assembled in this present parliament that peace be taken with England

10.3 Syncope

10.3.1 Plural nouns

The initial descriptive overview of syncope in npl {S} tokens showed that, overall, syncopated {S} is uncommon in the INFLAOS data. In terms of percentage per text, the occurrence of syncopated forms is high for certain outlying texts, whilst remaining very low on average. A potential reason for this distribution was suggested by plotting the percentage of syncopated npl {S} tokens by text grouped according to text type. This resulted in few outliers, and revealed that state documents contain the most syncope on average, but also that a significant amount of burgh records were also high in syncopated npl tokens. This distribution seems strange in light of the results presented in chapter 7, which set state documents and burgh records at opposite ends of the scale in terms of their palaeographical tendencies with regard to abbreviation. However, further investigation revealed that the distribution of syncopated forms across these two types of text is highly conditioned by the etymology of the lexel with which the syncopated form occurs. Specifically, the high level of syncope in burgh records is attributable to Germanic lexels, whereas the syncope in state documents occurs with non-Germanic lexels. This difference in which lexels are attested with syncopated inflections in these two types of text is mirrored by a difference in the orthographic realisation of the inflectional consonant. Syncopated Germanic {S} in burgh records is most often realised as <ȝ>, for example, *goods* in (41), whereas syncopated non-Germanic {S} in state documents most often appears as <s̓s̓>, for example, *imponitions* and *exactions* in (42). Having said that, a qualitative examination of the tokens in question showed that a large proportion of the <ȝ>-inflected tokens from burgh records can in fact be traced to the same manuscript and time-period, and furthermore mainly represent tokens of the same lexel, *good*. Of the 44 <ȝ>-inflected tokens attested in burgh records in the INFLAOS data, 41 come from the *Ayr Burgh Court Book* between 1444 and 1457. Of these 41, 31 are occurrences of <gudȝ> *goods*. {S} is also realised as <ȝ> in the same manuscript twice in <landȝ> *lands*, four times in <maist(er)ȝ> *masters* and four times in <p(ro)fetȝ> *profits*.

(41) <al gudʒ bocht wᵗout yᵉ towne>
all goods bought without the town [text 1104: 1452, BUR, AYR]

(42) <ye p(er)sonṡ punyst at ye kingſ wil>
the persons punished at the King's will [text 9510: 1471, STA, MLO]

The correlation between burgh records and the realisation of npl {S} as <ʒ> can therefore be attributed to the behaviour of the individual scribe or group of scribes who contributed to the *Ayr Burgh Court Book*. This is not to say that the occurrence of syncope in general is confined to particular scribes - only that the specific <ʒ> realisation is not attested elsewhere. On the contrary, syncope is shown by Model 4 (page 4) to be more likely in certain geographical areas, one such area being the south-west, around Ayr. It stands to reason that the area which stands out as displaying an unusual syncopated form is one where other syncopated inflection forms are most likely. If <gudʒ> and <landʒ> are variant forms of the more common <guds> and <lands> in which tailed z replaces <s>, then it makes sense for these variant forms to be employed by a scribe who is already likely to write {S} as <s>. On the other hand, there is the possibility that <ʒ> as a representation of {S} is not actually a tailed z representing a sibilant as the same *littera* clearly is in the word <kyndneʒ> *kindness* in (43), but rather an abbreviation symbol in the same way as <ſ>. This interpretation is favoured by Smith (2012), who notes his decision to transcribe separately the letters <ʒ> (tailed z, indicating a sibilant) and <ȝ> (yogh, indicating an approximant, as in <kyrkȝard> *kirkyard* in (44)), and states that he has "used ȝ [yogh] when it appears as a marker of plurality, e.g. seruauntȝ 'servants'" (Smith 2012: 71). This approach seems to suggest that Smith views final <ʒ/ȝ> in npl contexts as distinct from the tailed form of <z> used in (43). Smith does not elaborate on his decision to encode into his transcriptions a distinction between <ʒ/ȝ> as realisation of {S} and tailed z in other contexts, but there is evidence for this interpretation in the forms of abbreviation used in Medieval Latin. Cappelli (1982: 21) describes the evolution of an abbreviation symbol which began as a semi-colon and developed into "a mark like an arabic 3". This symbol could have several different meanings depending on the context in which it occurred, one of which being truncated *is* following a stem-final <s>. Cappelli only describes the semi-colon form as being used in this way, but does give an example, <ṙisꝫ> *remissis*, which shows the yogh-like form being used in place of <is>.

(43) <p(ro)cedand of luwe & kyndneʒ>
'proceeding of love and kindness' [text 368: 1450, CHA, PTH]

(44) yᵉ northest nok of yᵉ kyrkȝard
'the north east neuk [*corner*] of the kirkyard' [text 163: 1420, CAR, ROX]

There is no way to know whether the scribe of the *Ayr Burgh Court Book* was using <ʒ> as a letter z in its own right or as an abbreviation. He does not use this *littera* in sibilant contexts other than {S}, and often uses yogh in approximant contexts such as <ȝoung> *Young*, <ȝer> *year* and <balȝe> *bailie*. Having said

that, he also makes use of <ſ> to represent {S}, meaning that if his use of <ȝ/3> was an abbreviation, he would be using two abbreviation symbols for the same purpose. This, perhaps, seems intuitively unlikely, but for that matter, so does the incongruous use of tailed z to represent {S}.

As well as the effect of text type interacting with etymology, Model 4 indicated that the number of syllables in a token's stem was a significant predictor of the occurrence of a syncopated npl {S} inflection. The more syllables in the stem, the more likely the inflection is to be realised without a covered inflectional vowel, however, the significant difference in likelihood of syncope is between monosyllabic, disyllabic and trisyllabic stems, with little difference indicated between trisyllabic stems and stems with four or five syllables. This seems to indicate that the syllabicity of {S} is represented orthographically - as noted in chapter 2, the phonological process of weakening and eventual loss of the covered inflectional vowel began following trisyllabic stems, with monosyllabic stems retaining the syllabicity of the following inflection the longest.

If this is the case, however, it is surprising that the date of a text is not a significant predictor of inflectional syncope. Given that the covered inflectional vowel progressively weakened over time, culminating in its eventual complete loss, if the orthographically syncopated forms observed in INFLAOS mirror the phonological phenomenon, then it surely follows that the likelihood of orthographic syncope should increase over time. This is not evident in the INFLAOS data.

10.3.2 Past tense verbs & past participles

As with the equivalent npl data, an initial exploration of the percentage of syncopated vpt and vpp tokens in INFLAOS indicated an overall low level of {D} syncope. Also analogous to the distribution of npl {S} syncope is the fact that vpt and vpp syncope shows greater variation between individual texts than between individual lexels, suggesting that contextual factors, or even scribal choice may have had more of an influence on the occurrence of syncopated {D} forms than lexical factors.

The distribution of syncopated inflections in vpt and vpp tokens was also considered separately, revealing that on a textual level, more variation in the level of vpt syncope than vpp syncope is shown between individual texts. This may indicate that scribes were less systematic in their use of syncope with vpt tokens. A potential reason for this correlation may be related to contexts where there is ambiguity as to whether a past participle is in fact acting as a participial adjective. Fitting a GAM to the vpt and vpp INFLAOS data (Model 5; page 155) revealed an outlying lexel which illustrates this suggestion: *ken* 'know'. The vast majority of vpp *ken* tokens are realised as <kend> (as in (45)) or <kende> (as in (46)), with a small minority realised as <kennit> (as in (47)) or <kennyt> (as in (48)). In section 8.3.1.1, a qualitative analysis of the contexts in which vpp *ken* occurs showed that it is most often found at the beginning of a charter, as part of the set phrase *be it [made] kend til all men* 'be it made known to all men'. Examples of its use in this context are given in (45), (46) and (47). It is found only rarely elsewhere - an example of *ken* used outwith

this formula is given in (48). When used as part of the formulaic construction, 91% of vpp tokens of *ken* are realised as <kend> or <kende>.

(45) <Be it made kend tyll all me(n) be y(ir) p(rese)nt lett(er)is>
Be it made known to all men by these present letters [text 6: 1435, n.t., ELO]

(46) <Be it kende till meñ be thir(e) p(rese)nt l(ett)res>
Be it known to men by these present letters [text 26: 1490, CHA, n.l.]

(47) <Be it ke(n)nit till all men be y(ir) p(rese)nt l(ett)reȝ>
Be it known to all men by these present letters [text 232: 1438, CAR, FIF]

(48) <myselme(n)is wiſ᷑ acht to cu(m) to ye mar-kat wth cop & clapar y^{t} yai may be ke(n)nyt>
meselmen's [lepers'] *wives ought to come to the market with cup and clapper that they may be known* [text 1042: 1436, BUR, AYR]

10.3.3 Lexical Frequency and Formulaicity

The example of *ken* in section 10.3.2 raises the question of how lexical factors might impact the use of particular inflectional forms. Bybee (2007) describes the "lexical diffusion paradox", whereby "sound change seems to affect high-frequency words first, but analogical change affects low-frequency words first" (first observed by Schuchardt 1885 [1972]). The example she uses to illustrate this phenomenon with regard to analogical levelling is the analogical extension of the regular PDE <-ed> past-tense marker to low-frequency words in which, etymologically, pluralisation is denoted by medial-vowel umlaut, such as *leap - leapt*, which she finds are regularising: *leap - leaped*. Conversely, high-frequency words are more likely to retain the etymological umlaut plural: *sleep - slept*. Analogical extension is therefore more apparent in low-frequency tokens. The other side of the lexical diffusion paradox concerns sound changes, exemplified by Bybee with the example of schwa-reduction in PDE. Bybee finds that high-frequency words are more likely to exhibit schwa-deletion in spoken language than low-frequency words, a trend which Bybee attributes to "habituation". That is, the effect of a progressive sound change is most evident in words which are commonly used, with less common words retaining their pre-change phonetic values for longer.

If we assume that, on the evidence of (a) PDE reduction of the unstressed vowel in {D}; and (b) evidence from previous study (see section 2.2), that syncope of the unstressed vowel in {D} was a sound change in progress during the OSc period, the high occurrence of syncope in *ken* might be attributed to the high frequency of this word as part of a frequently-repeated formula used by scribes of legal documents. However, table 10.1 shows the 10 most frequently occurring lexels in INFLAOS by token count, along with the percentage of tokens which appear with a syncopated form of {D}. *ken* is the fifth most frequent lexel, and is syncopated in almost 90% of tokens, but most of the other most frequent vpp and vpt lexels do not show

the same high level of syncope. The most frequent vpp and vpt lexel is *call*, which is syncopated in only 5.46% of its 403 tokens. The exception to this is the second most frequent lexel, *oblige*, which appears in the syncopated form <oblist> 91.25% of the time in LAOS.

TABLE 10.1: The 10 most frequent vpt/vpp lexels in LAOS, excluding irregular forms. There are a total of 5,413 vpt/vpp tokens in INFLAOS. This table shows the 10 most frequently occurring lexels by token count, along with the percentage of tokens which appear with a syncopated form of {D}.

Lexel	# LAOS vpp/vpt tokens	Frequency rank	Syncope
call	403	1	5.46%
oblige	240	2	91.25%
ordain	233	3	14.59%
deliver	225	4	1.33%
ken	204	5	89.22%
compear	202	6	11.88%
accord	194	7	3.09%
ask	175	8	3.43%
grant	147	9	18.37%
contain	137	10	3.65%

The level of syncope in the most frequent vpt and vpp lexels in LAOS shown in table 10.1 suggests that the occurrence of syncope is not correlated simply with lexel frequency in the way that CIV reduction is correlated with word frequency in PDE.

Having said that, whilst frequency of individual lexels does not appear to correlate with syncope, the fact that *ken* occurs in a formulaic construction may itself influence its orthographic realisation. Kopaczyk (2013) describes OSc legal texts as containing "an impressive degree of structural repetition". These repeated structures encode "reference to the main legal activities of [...] burgh courts" and represent "fixed and stable patterns". Specifically, Kopaczyk refers to the repetition of fixed "lexical bundles" such as *be it made kend*.

Following this line of reasoning, the high occurrence of syncope in *ken* can be attributed to its status as a fixed form. Whereas scribes might ordinarily represent {D} as <it> or <yt> in the body of a text, the required formula to be repeated at the beginning of a specific type of document might not even be morphologically parsed - *be it made kend* was perhaps in the mind of a scribe, simply the required opening statement for a specific type of text.

There is one other lexel in table 10.1 which shows a high occurrence of syncope, *oblige*, realised often as <oblist>. Interestingly, this lexel is also identified by Kopaczyk (2013) as forming part of a formulaic bundle: *bind and oblige*. Indeed, an examination of the occurrences of vpp and vpt *oblige* in LAOS shows a large number of tokens occurring in this formula, often as <bundyn & oblist> or similar.

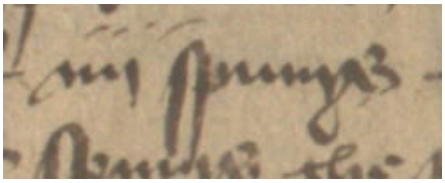
Whether there is a general tendency for formulaic bundles to exhibit syncopated forms is beyond the scope of this investigation, but it would make for interesting further analysis and potentially illumination of the factors influencing the occurrence of {D} syncope.

10.4 Covered inflectional <i> & <y>

10.4.1 Plural nouns

Section 9.2 investigated the alternation between <i> and <y> as the CIV in npl {S} tokens. The aim of this was to test the assumption that palaeographical factors conditioned which of these *littera* OSc scribes were likely to use. Specifically, the idea that the use of <y> as opposed to <i> was a method of increasing clarity in the context of a string of *littera* consisting of several minim strokes. Figure 10.1 illustrates the use of minim strokes with an example from text 1644 (1445, BUR, ABD) of the phrase <iiij spunys> 'four spoons'. The *litterae* <i>, <u> and <n> are formed using minim strokes, resulting in a consecutive stem-final row of four minims in the word *spoons*. This example is particularly apposite because it illustrates the tendency for scribes to realise the final *i* in a numeral string as <j> (i-longa), a practice inherited from Latin to prevent confusion with words formed of minim strings (Hector 1966: 41).

FIGURE 10.1: <iiij spunys> *four spoons* from text 1644 (1445, BUR, ABD).



The conclusion of the descriptive exploration and the GAM fitted to the data (Model 7; page 169) is that the use of <y> rather than <i> is conditioned by the number of consecutive minims in a token's SFL. Somewhat surprisingly, the correlation between the number of SFM and the likelihood of <y> as the CIV was observable only for SFM contained within the SFL. For example, a stem ending in four minim strokes forming the string <un>–such as <vexaciouniś> *vexations* 'physical violence' in (49)–does not appear to be more likely to be followed by covered inflectional <y> than a stem ending in two minim strokes forming a single letter <n>, such as <vexat(i)onis> in (50). Both of these <n>-final stems are more likely to be followed by covered inflectional <y> than non-minim-final stems, but it seems that the effect of minims in multiple consecutive stem-final *littera* does not have a cumulative effect on the likelihood of covered inflectional <y> as opposed to <i>.

(49) <ony sik yai sal happinė to sustenė be vexaciouniś>
any such they shal happen to sustain by vexations [physical violence] [text 433: 1473, CHA, STG]

(50) <costys & gret vexat(i)onis>
costs and great vexations [text 341: 1459, CAR, AGS]

Whilst the only significant fixed effect PV in Model 7 was SFM, the random effects of individual text

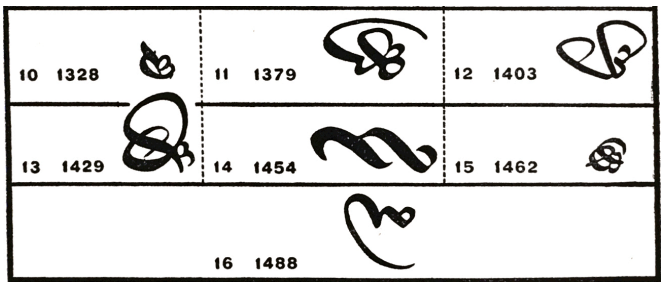
and lexel were also estimated to be significant in explaining the variation between <i> and <y> CIVs. Two lexels which stood out as unusually likely to occur with covered inflectional <y> were *mail* and *seal*. This recalls the observation made in chapter 7 that stem-final <l> could be realised in OSc handwriting with variant forms - a looped form and a straight form. In the case of {S} abbreviation, it was suggested that the somewhat unpredictable distribution of <ſ> following <l>-final stems could be due to these variant forms. The looped form, with its horizontal final stroke, would flow smoothly into a following <ſ>, similarly to the looped ascender of <d>. The straight form, finishing in a vertical down-stroke, would not flow into <ſ>, much like the final minim-stroke of <m> or <n>. In the case of covered inflectional <y>, this distribution might have a similar effect in that a vertical <l> might be similar enough to an extended minim-stroke to create the same <i>-repelling effect.

At the other end of the likelihood scale, two lexels in particular suggest another factor which might have an effect on whether a scribe used <i> or <y> as the CIV: *bigging* 'building' and *heir* are both especially unlikely to occur with covered inflectional <y>. A potential reason for this might be that both of these lexels occur very often with stem-medial <y>, as <ayr-> and <byggyng-> respectively. In particular, *bigging* has the potential to be realised not only with two medial <y>s, but also three medial <g>s, as in (51). Whilst there is not the same risk of *litteral* ambiguity in a string of <y>s and <g>s as with a string of minims, it is still possible that scribes chose to avoid too many repetitions of similar letter-shapes. In the case of <y> and <g>, the presence of too many curling descenders might have been avoided in the interest of readability, particularly if lines of script were written close together.

(51) <oñ ye west sid of yis byggyng yar is a dik of stan & thorn>
on the west side of this bigging [building] *, there is a dyke of stone and thorn* [text 205: 1435, CHA, FIF]

Two texts which showed atypical distributions of covered inflectional <y> were already identified at the descriptive stage of analysis. Texts 63 and 131 stood out by virtue of being the only two state documents which contain a high percentage of npl {S} inflections realised as <ys>. Interestingly, npl covered inflectional <y> in both of these texts is confined almost entirely to inflected forms of a single lexel. In text 131, the word *trews* 'truce' is realised eight times as <trewys> and once as <trewis>. There is also one instance of <realmys> *realms* which conforms to the expected model of covered inflectional <y> following stem-final <m>, a *littera* formed of three minim strokes. Covered inflectional <y> following stem-final <w>, however, is not predicted by this generalisation, as <w> is not typically written with minim strokes, but with curved pen-strokes, as shown in the examples in figure 10.2, reproduced from Johnson and Jenkinson (1915: 52). Furthermore, the scribe uses <i> as the CIV in a vpt {D} inflection following stem-final <w> elsewhere in the text: <sewit> *sued*.

FIGURE 10.2: Potential realisations of <w> between 1328 and 1488, reproduced from Johnson and Jenkinson (1915: 52)



A possible explanation for the unusual use of <ys> in text 131 is the fact that the plural form *trews* underwent a process of reanalysis as a singular form, becoming *truce* in ModSc and PDE, with the plural form *truces*. According to Bosworth (1898a: 1013), the usage of the plural form of *treōw* 'truth, faith' (*treōwa*) with singular meaning was already evident in OE. Singular *trew* is attested in LAOS, as shown in (52), indicating that the transition to reanalysed singular *truce* was not entirely complete, but the singular form is much less frequent than the plural, with 10 tokens of *trew* attested in LAOS compared to 72 tokens of *trews*. It may be that the scribe of text 131 considered *trews* to be a singular form rather than an inflected plural noun. In that case, there is no real reason to expect the final segment of *trews* to match the scribe's usual representation of {S}.

(52) <ye saide co(m)misairſ has takin a trewe for a yhere>
the said commissars [delegates] *have taken a truce for a year* [text 130: 1409, STA, ROX]

10.4.2 Past tense verbs & past participles

Given that the investigation into the use of covered inflectional <y> in npl {S} strongly supported the idea that palaeographical context is the crucial factor, it seems reasonable to expect that the same would be true for vpt and vpp {D} inflections. However, the GAM presented in section 9.3.1 suggested that several other factors correlate significantly with the use of covered inflectional <y> in {D}. In particular, text location was found to predict the likelihood of <y>. A heatmap of the likelihood of <y> overlaid on a geographical map of Scotland identified specific areas where <y> is particularly prevalent. Using this information, the percentage of <y> in tokens with different SFL was plotted for the four counties with the most data points. This showed a clear difference between the distribution of <y> in Fife, an area in which covered inflectional <y> was identified as likely, and three other counties where <y> was identified as unlikely (Aberdeen, Ayr and Midlothian). In texts localised to one of the latter three counties, the percentage of vpt and vpp tokens with covered inflectional <y> is generally low, but markedly higher following stem-final <m>, indicating conditioning of the variant by SFM count. In Fife, on the other hand, the difference between the percentage of <y> tokens for each SFL is noticeably smaller. Stem-final <m> still has the highest percentage, but there

is much less of a <y>/<i> divide between <m> and the other, non-minim SFL than in the other counties. This suggests that the use of covered inflectional <y> was conditioned to some degree by scribal preference. The fact that Fife as a county stands out from the surrounding geographical area suggests that the preference for <y> over <i> may have constituted a stylistic choice by a group of scribes.

10.5 Covered inflectional <e>

The use of <e> as the CIV is not as straightforward to investigate as the use of <i> or <y>. This is because of the ambiguity of status of covered <e> as part of the stem or part of the inflection. Stem-final <e> in singular nouns is extremely prevalent in the LAOS data and, as discussed in section 9.4.1, there are a number of potential factors conditioning its presence or absence. The investigation into <e> therefore began with a comparison of the occurrence of <e> in plural and in singular noun tokens. This comparison focussed on a subset of lexels, chosen based on their having a sufficient number of tokens attested in both singular and plural forms. Restricting the dataset in this way ensured that the results of the comparison would not be biased by particular lexels which occur often in their plural form but rarely in the singular, or vice versa. For example, *earl* occurs 448 times in the singular in INFLAOS, whereas in the plural it occurs only three times. The subset of lexels was also restricted to monosyllabic words, to avoid the complicating factor of syllable stress and enable the classification of lexels based on the length of their medial vowel (based on the system of OSc vowels proposed by Aitken and Macafee (2002)).

The results of a GAM fitted to the npl data subset (Model 10; page 182) indicated that the likelihood of covered <e> in plural nouns can be predicted with a significant level of accuracy by a combination of contextual and lexical factors. Notarial protocol books are most likely to contain covered <e>, and cartularies least likely. The date as well as the type of a text is a significant predictor of covered <e> in npl tokens. Two predictors showed a significant interaction with date: etymology and medial vowel length. The likelihood of covered <e> tokens of non-Germanic lexels appears to have remained fairly consistent over time, whereas the likelihood of the same phenomenon in Germanic tokens fluctuated. In particular, there is a steep drop in the likelihood of covered <e> in Germanic plural nouns in the period 1380-1440. This drop is suggestive of the trend described by King (1997), whereby an initial high level of <e> attributable to English scribal influence quickly gave way to the characteristically Scots <i>/<y> CIVs. King's explanation does not account for the difference in covered <e> occurrence between Germanic and non-Germanic tokens if the use of covered <e> in OSc is attributed entirely to fleeting English influence. However, it is possible that English scribal influence contributed more covered <e> in Germanic tokens than would otherwise have occurred in OSc, but that the use of covered <e> in non-Germanic tokens was similar in OSc and ME. In this case, the decline of the early ME scribal influence could manifest itself as a decrease in Germanic covered <e>, whilst

non-Germanic covered <e> continued to be used by OSc scribes.

Etymology was also shown to interact with date in predicting final <e> in singular nouns, but the correlation indicated is different to that which characterises the likelihood of covered <e> in plural nouns. Non-Germanic tokens do not show any meaningful change with regard to the likelihood of final <e> across the period 1380-1500, but for Germanic tokens, the likelihood of <e> appears to gradually decrease between 1380 and 1470, before experiencing a sharper decline in likelihood in the final 30 years of the fifteenth century. The difference between the trajectory of <e> likelihood over time for singular and for plural nouns suggests that final and covered <e> do not show the same patterns of attestation and therefore that they are separate phenomena. More specifically, if the covered <e> observed in npl tokens is in fact the same <e> observed in final position in singular nouns, then a decline in the likelihood of final <e> should be mirrored by a corresponding decline in the likelihood of covered <e>, which is not the case.

Medial vowel length was also shown to interact with date in Model 10. In this case, the difference between the likelihood of singular final <e> and of npl covered <e> over time is even more apparent. Singular final <e> is significantly more likely to occur after a long medial vowel, suggesting that it was employed as a length diacritic. However, as time goes on, the gap between the likelihood of final <e> following long and short medial vowels appears to narrow substantially, suggesting an increasing use of 'otiose' final <e>. Conversely, the likelihood of covered <e> following long and short medial vowels remains similar for most of the period, though is an apparent divergence in the second half of the fifteenth century, whereby covered <e> increases in likelihood following long medial vowels. There is no real evidence to link the temporal distributions of final and covered <e> in terms of their occurrence after long and short medial vowels, but a speculative suggestion might be that the increased likelihood of covered <e> following long medial vowels is a belated application of <e> as a stem-final length diacritic in plural nouns, even as its use begins to decline in the singular.

10.6 Statistics and historical dialectology

Chapter 1 introduced the idea of using regression modelling techniques to capitalise on the recent increase in sources of 'big data' for historical linguistics. Throughout the investigation in the following chapters, this methodology has been applied. Fitting GAMs with random effects structures as advocated by Gries (2015: 97) allowed me to take into account the hierarchical structure of the LAOS data. For all the DVs investigated as part of this survey of the inflectional orthography of OSc, the GAM results showed that there was significant variation in the data due to the individual differences between texts or between lexels. In most cases, both of these sources of random variation were able to contribute significantly to understanding the distribution of the data. The inclusion of these effects allowed the main effects, the variables of interest

in each model, to be interpreted with confidence that any significant correlation indicated was not the result of a few 'rogue' scribes or lexels. Furthermore, where individual texts and lexels showed an unexpected distribution with regard to a DV, the GAM was able to identify these outliers. This in turn allowed for detailed qualitative analysis of potential reasons behind these unexpected distributions. In many cases, simply knowing which texts or lexels to zoom in on in detail revealed factors affecting the realisation of inflections in particular contexts.

10.6.1 The need for qualitative analysis

Throughout Chapters 6 to 9, complex and broad statistical analyses of the INFLAOS data have been presented alongside more detailed, qualitative insights. This is a practice which is especially necessary when analysing Medieval manuscript data from sources as varied, both in terms of spatio-temporal context and subject matter, as those included in LAOS.

At first glance, it may seem that complex statistical analysis of large corpora is the antithesis of detailed qualitative analysis. I hope to have shown, however, not only that these two approaches are compatible, but that high-level modelling of such data is instrumental in revealing avenues for more traditional qualitative analysis. This is not to say, however, that the benefit is one-sided - qualitative analysis is not only aided and focussed by broad statistical analysis, it is often necessary to make sense of the insights such analysis provides. A clear example of this is the significance of SFL in the likelihood of abbreviation. The GAM presented in section 7.2.1 showed clearly that the PV SFL can be divided into two groups: a group with high likelihood of <ſ> and a group with low likelihood of <ſ>. However, the link between the SFL which make up these two groups is clear only on inspection of the manuscript forms themselves. To make sense of the fact that, for example, <m> is far less likely to be followed by <ſ> than <d>, it is necessary to see the palaeographic forms of the letters and make a logical deduction about the physical formation of them by a scribe. This is an aspect of analysing medieval manuscripts which does not apply to studies in which the primary data source is digital, and which have led the way in the application of regression modelling techniques to corpus data. Nonetheless, these techniques allow the researcher of medieval manuscripts to simultaneously examine the correlation of many contextual and lexical factors with transcribed manuscript forms. This has the advantage, not only of revealing whether specific predictor variables significantly correlate with particular features when all PVs are taken into account, but also of controlling for the variance that idiosyncratic features of particular texts or words cause in such a dataset. Some examples of this kind of variance which have been discussed in this chapter are the occurrence of npl zero for tokens of *pertinents* in the notarial protocol books of Thomas Crawfurd; and the occurrence of vps zero for tokens of *will* in the formulaic construction *as the law will*.

10.6.2 A limitation of the methodology

A source of variation which is inaccessible even to the most detailed qualitative analysis is the identity of the individual scribe who penned each LAOS text. In all the GAMs fit to the INFLAOS data during this investigation, the random effect of individual text has been investigated and used as a kind of proxy for scribal variation. In conjunction with the fixed effects of date and location, this is as close as it is possible to come to modelling the correlation of DVs with individual scribes. That is, under the assumption that scribes were typically active in a particular location at a particular time and did not regularly travel great distances to produce large numbers of manuscripts in far-flung areas, the predictors of text date and location should capture this variation indirectly, by indicating the presence of dates and locations which have an unusually high or low likelihood of the orthographic form under investigation. If a particular scribe stands out from his spatio-temporal peer group, then the inclusion of the random effect of individual text in the model will reveal this by indicating an unusually high likelihood of the form in the texts attributable to that scribe. However, without access to the specific identity of the scribe of each text, it is not possible to know whether one scribe or several are contributing to an observed trend in a particular location at a particular time.

Chapter 11

Conclusion

11.1 The distribution of inflectional forms in Older Scots (OSc)

The aim of this thesis was to investigate, describe and analyse the orthographic representations of {S} and {D} inflections in OSc using the *A Linguistic Atlas of Older Scots* (LAOS) data. Chapter 2 provided an overview of the extant scholarship on this topic and demonstrated that it has been largely confined to (a) broad assumptions about the 'typical' representation of inflections in OSc manuscripts, often conflating full and abbreviated representations of {S}; and (b) small data-driven investigations which have been severely limited in scope by the scant availability of transcribed OSc material which preserves the reality of written manuscripts in sufficient detail to allow thorough analysis.

Chapter 3 identified the main questions arising from the review of the literature in chapter 2. The research questions identified the status of abbreviated forms of plural noun (npl) and present tense verb (vps) {S} as a particular area of interest. In particular, whether it was used entirely interchangeably with fully-realised forms of {S} by OSc scribes, or whether there are other factors conditioning its attestation in OSc texts. Another issue relating to the specific forms used by OSc scribes which, like abbreviation, is often subject to conflation with the 'typical' covered inflectional <i> realisation in transcribed texts, is the alternation between <i> and <y>. The use of covered inflectional <e> in OSc manuscripts is also an area of interest, particularly in light of the account given by King (1997) of the temporal variation in its use in OSc and its relation to Middle English (ME) scribal influence.

Chapter 4 introduced LAOS as the solution to the previous paucity of detailed transcriptions of osc texts, and outlined the extraction from LAOS of a dataset of {S} and {D} inflectional forms, dubbed *Inflections in A Linguistic Atlas of Older Scots* (INFLAOS). The amount of data in INFLAOS was sufficient to not only provide an accurate description of the distribution of the orthographic forms of these inflections, but also to

apply an innovative and robust statistical analytic methodology to investigate the conditioning factors which correlate with the use of particular orthographic variants. This methodology was explained and exemplified in chapter 5, where it was shown that it is necessary to account for the hierarchical structure of corpus data such as LAOS in order to reliably identify the variables which correlate with the dependent variable (DV) under investigation. As outlined in chapter 1, the 'opportunistic' nature of the data included in LAOS makes analysis of it particularly susceptible to influence by outlying texts or lexels which show atypical forms.

Chapter 6 presented an analysis of the use of zero-morphemes in INFLAOS. The distinction between zero and non-zero realisations of past tense verb (vpt) and past participle (vpp) {D} was shown to correspond to the distinction between regular and irregular lexels, but zero inflection in {S} tokens presented a more complex picture. The stem-final string of npl tokens was shown to correlate with the use of zero if it was similar in form to the fully-realised form of {S}. Stem-final <s> alone did not produce the same effect, but the occurrence of <ß> as the stem-final *littera* (SFL) was shown to significantly predict a zero realisation of {S}. In vps tokens, the use of zero was strongly correlated with the operation of the Northern Subject Rule (NSR), though stem-final <ß> was still shown to predict a zero realisation in non-NSR environments, suggesting a general avoidance of the use of fully-realised {S} following stem-final <ß> by OSc scribes. The use of zero in npl {S} tokens was also shown to vary significantly over time and space, as well as being mor e likely in certain types of text. These contextual factors were not significant in the realisation of vps zero, presumably due to its status as a codified part of the OSc verbal paradigm.

Chapter 7 investigated the use of the abbreviation <ſ> to represent {S}, finding a clear explanation for its distribution in the shapes of the *littera* after which it occurred. A final horizontal stroke extending rightwards from the SFL in a token was shown to be the crucial factor determining whether a scribe would use the abbreviated form. Having said this, the use of abbreviation also increased over time, potentially due to the increasing use of OSc as opposed to Latin in legal documents which led to scribes increasing development and use of abbreviation as a time-saving strategy. Certain text types were also more likely to contain abbreviation than others. The high attestation of <ſ> in burgh records compared to the low attestation in state documents suggests that more important and formal documents written by professionally trained scribes contained fewer abbreviated inflections and more fully-realised forms.

Chapter 8 analysed the use of {S} and {D} forms with no covered inflectional vowel in INFLAOS. Both lexical and contextual factors correlated with the use of syncopated forms of {S}. Specifically, non-Germanic words with more than one syllable were most likely to be realised as a syncopated form, particularly if the token occurred in a state document. In the case of {D} inflections, a distinction was shown between the attestation of vpp and vpt syncope, with the contextual factors of text type and location influencing the use of syncopated forms in vpp but not in vpt tokens. The occurrence of stem-final <s> and <x> also predicted the syncope of vpp {D}, but this correlation was not mirrored by vpt {D}. The model fit to the data for vpp

explained more of the variation between tokens with regard to syncope, whereas the model fit to the vpt data showed that there was still a large amount of random variation unaccounted for.

Chapter 9 examined the possible realisations of the covered inflectional vowel (CIV) in INFLAOS, finding significant support for the hypothesis that <y> was used instead of <i> to disambiguate strings of minim-strokes in {S} inflections. However, the same was not true for realisations of {D}, in which the use of <y> was shown to be potentially influenced by scribal choice rather than necessity. Crefch:civ also considered the use in OSc of covered inflectional <e>, more generally associated with ME. The difficulty of separating the phenomena of stem-final and covered inflectional <e> was acknowledged, and an investigation performed to see whether there was evidence to interpret <e> in covered inflectional position as a part of the stem rather than the inflection. A comparison of models fit to singular and plural noun data from LAOS suggested that stem-final and covered inflectional <e> did not follow the same patterns with regard to the contextual and lexical factors which conditioned them.

Finally, chapter 10 gave a short recap of the results of the investigations into each inflectional realisation, interpreting the results in light of both the statistical analyses and the qualitative investigations prompted by them. The end result of this thesis is a detailed overview of the various orthographic representations used by OSc scribes to denote {S} and {D} inflections, accompanied by various focussed analyses of individual texts, scribes, manuscripts and lexels.

11.2 Written texts as evidence of linguistic change

The subject of this thesis was conceived as part of the ongoing *From Inglis to Scots* (FITS) project (Kopaczyk et al. 2018) at the Angus McIntosh Centre for Historical Linguistics at the University of Edinburgh. The purpose of FITS is to "elucidate the underlying sound system [of OSc], via the orthographic alternations within the Germanic morphemes of the [LAOS] corpus" (Molineaux 2014).

The method used in FITS to analyse the underlying phonology of the LAOS texts is "grapho-phonological parsing" (Kopaczyk et al. 2018). Each token of a given Germanic lexel is deconstructed into its component *littera*, with each *littera* or group of *litterae* representing one or more potential sound values. The sound value attributed to the *littera* or *litterae* is determined based on several factors, described by Kopaczyk et al. (2018):

(a) "spelling variation across the corpus";

(b) "what we know about the sounds of mediaeval Scots (e.g. based on Aitken and Macafee 2002 and Johnston 1997)";

(c) "what we know about the sounds of the preceding and following stages of the language"; and

(d) "general theories of sound change and language change."

Applying this methodology to the inflectional morphemes in LAOS does not provide very much evidence to support a definitive attribution of underlying phonetic values. As described in the review of the literature in chapter 2, the unstressed vowels of OSc have not previously been investigated in detail, with the most detailed analyses focussing on verse evidence as opposed to the phonetic implications of inflectional realisations in prose texts.

The evidence provided by preceding and following stages of the language is available however - we know that the inflectional vowel underwent a process of reduction and ultimately loss between Œand Present Day English (PDE). Given the high occurrence of <i> and <y> as the inflectional vowel in OSc {S} and {D}, it seems safe to assume that in the formative stages of the development of OSc from various varieties of ME, the unstressed vowel was present to a sufficient degree that its raised realisation was manifested in the orthographic representation of {S} and {D}.

As demonstrated in the analysis of {S} and {D} syncope in chapter 8, OSc texts do not show a diachronic increase in syncope over the period 1380-1500. Such a trend would suggest reduction and loss of the inflectional vowel over time reflected in the orthography, but the absence of direct orthographic evidence does not necessarily imply the absence of this process. Orthographic representations often persist long after a sound change has occurred, for example, the <gh> in PDE *night*, indicating a historical fricative. Of course, standardisation of spelling plays a large part in this phenomenon in PDE, a process to which OSc spelling was not subject. However, despite the absence of spelling standardisation in the sense in which it applies to PDE, there was a form of textual standardisation present in the LAOS texts by virtue of their status as legal documents. As described by Kopaczyk (2013), the scribes of these texts constituted "a special type of a professional community who had been trained to conform to specific discourse rules and traditions". Because of the importance of consistency in the language of legal record-keeping, scribes made use of what Kopaczyk describes as "fixed and stable patterns". Whilst Kopaczyk's study refers to lexical formulae, it must be recognised that the texts contained in LAOS are subject to constraints of formulaicity, and that this may have implications for their use of inflectional forms. Specifically, scribes of legal documents may have been more likely to retain conventions of the written language even if those conventions no longer accurately represented the spoken language.

Bibliography

Ackermann, A. (1897). "Die Sprache der ältesten schottischen Urkunden, A.D. 1385-1440." PhD thesis. Georg-Augusts Universität zu Göttingen.

Aitken, A. J. (1977). "How to Pronounce Older Scots". In: *Bards and Makars: Scottish Language and Literature Medieval and Renaissance*. Ed. by A. J. Aitken, M. P. McDiarmid, and D. S. Thomson. Glasgow: University of Glasgow Press, pp. 1–21.

— (1985). "Introduction. A history of Scots". In: *The concise Scots dictionary*. Ed. by M. Robinson. Repr. [Aberdeen]: Aberdeen University Press, 819p.

— (1997). "The Pioneers of Anglicised Speech in Scotland: a second look". In: *SCOTTISH LANGUAGE* 16, pp. 1–36.

Aitken, A. J. and C. Macafee (2002). *The Older Scots Vowels: A History of the Stressed Vowels of Older Scots from the Beginnings to the Eighteenth Century*. Edinburgh: Scottish Text Society.

Archibald, E. P. (2013). "Whose Line is it Anyway? Dialogue with Donatus in Late Antique and Early Medieval Schools". eng. In: *The Journal of Medieval Latin* 23, pp. 185–199.

Aw v. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Awe v.1 (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bann, J. and J. Corbett (2015). *Spelling Scots: the orthography of literary Scots, 1700-2000*. Edinburgh: Edinburgh University Press.

Beal, J. (1997). "Syntax and morphology". In: *The Edinburgh History of the Scots Language*. Ed. by C. Jones. Edinburgh: Edinburgh University Press, pp. 335–377.

Bosworth, J. (1898a). *An Anglo-Saxon Dictionary: Based on the Manuscript Collections of the Late Joseph Bosworth.* Ed. by T. N. Toller. Oxford: Clarendon.

— (1898b). *Dwellan. Based on the Manuscript Collections of the Late Joseph Bosworth*. In: *An Anglo-Saxon Dictionary*. Ed. by T. N. Toller. Oxford: Clarendon, p. 220.

Bound n.1 (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Breheny, P. and W. Burchett (2016). *visreg: Visualization of Regression Models*. R package version 2.3-0.

Bugaj, J. (2002). "Verb morphology of South-Western Middle Scots". In: *Studia Anglica Posnaniensia: international review of English Studies*, p. 49.

— (2004a). *Middle Scots inflectional system in the south-west of Scotland*. Frankfurt am Main: Peter Lang.

— (2004b). "'for ye vrangus haldyn of thre bollis of beire fra hyre': Nominal plurals in south-western Middle Scots". In: *Linguistica e Filologia* 19, pp. 53–74.

Burgeis, n. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Burgen, n. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Cappelli, A. (1982). *The Elements of Abbreviation in Medieval Latin Paleography*. Trans. by R. Heimann David; Kay. Lawrence: University of Kansas Libraries.

Dampnis n. pl. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Dieth, E. (1932). *A Grammar of the Buchan Dialect*. Cambridge.

Disassent, v. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Grant, W. and J. M. Dixon (1921). *Manual of Modern Scots*. Cambridge: Cambridge University Press.
Gries, S. (2015). "The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models". In: *Corpora* 10.1, pp. 95–125.
Gries, S. T. (2003). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. Open Linguistics. Bloomsbury Publishing.
Haas, N. K. de (2011). *Morphosyntactic Variation in Northern English: The Norhern Subject Rule, its Origins and Early History*. Nijmegen: Radboud University.
Haas, N. de and A. Van Kemenade (2015). "The origin of the Northern Subject Rule: subject positions and verbal morphosyntax in older English". In: 19.1, pp. 49–81.
Haugen, E. (1950). "First Grammatical Treatise. The Earliest Germanic Phonology". In: *Language* 26.4, pp. 4–64.
Hector, L. C. (1966). *The Handwriting of English Documents*. London: Edward Arnold.
Hogg, R. M. (1992). "Phonology and morphology". In: *The Cambridge History of the English Language*. Ed. by R. M. Hogg. Vol. 1. Cambridge University Press, pp. 67–167.
Jenkinson, H. (1937). *A Manual of Archive Administration*. London: P. Lund, Humphries & Co. Ltd.
Johnson, C. and H. Jenkinson (1915). *English Court Hand, A.D. 1066 to 1500*. Vol. 2. Oxford: Clarendon.
Johnston, P. (1997). "Older Scots Phonology and its Regional Variation". In: *The Edinburgh history of the Scots language*. Ed. by C. Jones. Edinburgh: Edinburgh University Press, pp. 47–111.
Jordan, R. and E. J. Crook (1974). *Handbook of Middle English Grammar: Phonology*. The Hague: Mouton de Gruyter.
King, A. (1997). "The Inflectional Morphology of Older Scots". In: *The Edinburgh History of the Scots Language*. Ed. by C. Jones. Edinburgh: Edinburgh University Press, pp. 156–181.
Kniezsa, V. (1997). "The Origins of Scots Orthography". In: *The Edinburgh history of the Scots language*. Ed. by C. Jones. Edinburgh: Edinburgh University Press, pp. 24–46.
Kopaczyk, J. (2001). "The Scots-Northern English continuum of marking noun plurality". In: *Studia Anglica Posnaniensia: international review of English Studies*, p. 131.
— (2013). *The Legal Language of Scottish Burghs: Standardization and Lexical Bundles (1380-1560)*. Oxford studies in language and law. Oxford: Oxford University Press, [2013].
Kopaczyk, J., B. Molineaux, V. Karaiskos, R. Alcorn, B. Los, and W. Maguire (2018). "Towards a graphophonologically parsed corpus of medieval Scots: database design and technical solutions". eng. In: *Corpora* 13.2, pp. 255–269.
Kuipers, C. (1964). *Quintin Kennedy, 1520-1564: Two Eucharistic Tracts*. Nijmegen: Drukkerij Gebr. Janssen.
Laing, M. (1999). "Confusion wrs Confounded: Litteral Substitution Sets in Early Middle English Writing Systems". In: *Neuphilologische Mitteilungen: Bulletin de la Société Néophilologique* 100.3, pp. 251–70.
— (2004). "Multidimensionality: Time, Space and Stratigraphy". In: *Methods and data in English historical dialectology*. Ed. by M. Dossena and R. Lass. Bern ; Oxford: P. Lang. Chap. Linguistic insights: studies in language and communication v. 16, 405 p. ; 23 cm.
— (2009). "Orthographic indications of weakness in Early Middle English". In: *Phonological Weakness in English: from Old to Present-Day English*. Ed. by D. Minkova. New York: Palgrave Macmillan, pp. 237–315.
— (2013). *A Linguistic Atlas of Early Middle English, 1150-1325*. Version 3.2. URL: `lel.ed.ac.uk/ihd/laeme2/laeme2.html`.
Laing, M. and R. Lass (2003). "Tales of the 1001 Nists: The Phonological Implications of Litteral Substitution Sets in Some Thirteenth-Century South-West Midland Texts". In: *English Language and Linguistics* 7.2, pp. 257–278.
— (2009). "Shape-shifting, sound-change and the genesis of prodigal writing systems". In: *English Language and Linguistics* 13.1, pp. 1–31.
— (2013a). "Introduction, part I: background". In: *A Linguistic Atlas of Early Middle English*. Chap. 2.
— (2013b). "Introduction, part II: the corpus". In: *A Linguistic Atlas of Early Middle English*. Chap. 3.
Lass, R. (1976). *English Phonology and Phonological Theory: Synchronic and Diachronic Studies*. New York: Cambridge University Press.

— (1992). "Phonology and morphology". In: *The Cambridge History of the English Language*. Ed. by N. Blake. Vol. II, pp. 23–155.
— (2009). "On schwa: synchronic prelude and historical fugue". In: *Phonological Weakness in English: from Old to Present-Day English*. Ed. by D. Minkova. New York: Palgrave Macmillan, pp. 47–77.
Lass, R., M. Laing, R. Alcorn, and K. Williamson (2013). *A Corpus of Narrative Etymologies from Proto-Old English to Early Middle English and accompanying Corpus of Changes*. Version 1.1. The University of Edinburgh. URL: http://www.lel.ed.ac.uk/ihd/CoNE/CoNE.html.
Livy ([1533] 1901). *Livy's History of Rome, The First Five Books. Translated into Scots by John Bellenden 1533*. Ed. by W. Craigie. Trans. by J. Bellenden. Edinburgh: Blackwood.
Macafee, C. (1983). *Glasgow*. Vol. 3. Varieties of English Around the World. Amsterdam: Benjamins.
— (2002). "Introduction". In: *The Older Scots Vowels: A History of the Stressed Vowels of Older Scots from the Beginnings to the Eighteenth Century*. Edinburgh: Scottish Text Society.
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". PhD thesis. The University of Edinburgh.
McIntosh, A., M. L. Samuels, and M. Benskin (1986). *A Linguistic Atlas of Late Mediaeval English*. Aberdeen: Aberdeen University Press.
Minkova, D. (1991). *The History of Final Vowels in English: The Sound of Muting*. Berlin: Mouton de Gruyter.
— (2014). *A Historical Phonology of English*. Edinburgh textbooks on the English language. Advanced. Edinburgh: Edinburgh University Press.
Molineaux, B. (2014). *From Inglis to Scots: Mapping Sounds to Spellings (FITS)*. URL: http://www.amc.lel.ed.ac.uk/?page_id=498.
Month n.1 (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.
Murray, J. A. H. (1873). *The Dialect of the Southern Counties of Scotland*. London: Asher & Co.
Ordnance Survey Ltd. (2016). *Batch coordinate transformation tool*. URL: https://www.ordnancesurvey.co.uk/gps/transformation/batch.
Oxford English Dictionary (2018). *truce, n.* In: *OED Online*. Oxford University Press.
Pertin-, Pertenence, -ens n. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.
Pertinent n. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.
Petrović, S., M. Osborne, and V. Lavrenko (2010). "The Edinburgh Twitter Corpus". In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. WSA '10. Los Angeles, California: Association for Computational Linguistics.
Prot(h)ogoll, -col(l, n. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.
Rodriguez Ledesma, M. (2013). "The Northern Subject Rule in first-person singular contexts in fourteenth-fifteenth-century Scots". In: *Folia Linguistica* 34, pp. 149–172.
Romaine, S. (1982). "The English Language in Scotland". In: *English as a World Language*. Ed. by R. W. Bailey and M. Görlach. Ann Arbor: University of Michigan, pp. 56–83.
Schuchardt, H. (1885 [1972]). "On sound laws: Against the Neogrammarians". In: *Schuchardt, the Neogrammarians, and the transformational theory of phonological change*. Ed. by T. Venneman and T. H. Wilbur. Frankfurt am Main: Athenaum.
Sharpsteen, C. and C. Bracken (2016). *tikzDevice: R Graphics Output in LaTeX Format*. R package version 0.10-1.
Simpson, G. (1973). *Scottish Handwriting, 1150-1650: An Introduction to the Reading of Documents*. Aberdeen: Aberdeen University Press.
Smith, D. (2019). "The Predictability of {S} Abbreviation in Older Scots Manuscripts According to Stem-final *Littera*". In: *Historical dialectology in the digital age*. Ed. by R. Alcorn, J. Kopaczyk, B. Los, and B. Molineaux. Edinburgh: Edinburgh University Press.
Smith, J. J. (2012). *Older Scots: A Linguistic Reader*. Edinburgh: Scottish Text Society, xi, 253 pages ; 24 cm.

Speelman, D (2014). "Logistic regression: A confirmatory technique for comparisons in corpus linguistics". In: *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*. Ed. by D. Glynn and J. Robinson. Amsterdam: John Benjamins, pp. 487–533.

Strang, B. (1970). *A History of English*. London: Methuen.

Trew(is n. (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wieling, M. and J. Nerbonne (2015). "Advances in Dialectometry". In: *Annual Review of Linguistics* 1.1, pp. 243–264.

Wieling, M., S. Montemagni, J. Nerbonne, and R. H. Baayen (2014). "Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling". In: *Language* 90.3, pp. 669–692.

Wil(l, v.1 (2004). In: *Dictionary of the Scots Language*. Scottish Language Dictionaries Ltd.

Williamson, K. (2008). *A Linguistic Atlas of Older Scots*. Version 1.2. URL: `http://www.lel.ed.ac.uk/ihd/laos1/laos1Z.html`.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman & Hall/CRC.

Wyntoun, A. ([1350-1420] 1872-79). *The orygynale cronykil of Scotland*. Ed. by D. Laing. Edmonston and Douglas. Vol. I. Edinburgh.

Zuur, A. F. (2009). *Mixed effects models and extensions in ecology with R*. New York: Springer.

Appendix A

Frequency of tokens omitted from INFLAOS

Reason for omission		Tokens omitted	Tokens remaining
			38,152
Vowel-final lexels including $do and $be		8,077	30,075
Specific lexels	$depute	5	30,070
	$have	1,778	28,292
Specific texts	T88	95	28,197
	T97	2	28,195
	T160	11	28,184
	T8001	34	28,150
Forms including [?]	Stem-final [?]	102	28,048
	Inflectional [?]	7	28,041
Rehiving <(-)e(-)>	$letter	801	27,240
Vc strings	npl abbr. = <(er)>	1	27,239

Reason for omission		Tokens omitted	Tokens remaining
	npl abbr. = <(us)>	1	27,238
	Low-frequency <VC> tokens	61	27,177
Specific lexels	Plural lexels: $matins & $stocks	8	27,169
	Onomastic lexels: $blackfriars & $peeble	2	27,167
SFL dataset reduction - infrequent categories	Superscript SFL	150	27,017
	SFL = low-frequency abbr.	51	26,966
Final total			**26,966**