Chapter Five

Uncovering linguistic lineage through using a character N-gram based dialect classifier

Kevin Buckley

University of Newcastle

Abstract

Quantitative approaches to analysing diachronic change have become popular in examining historical languages (Piotrowski 2012). Character N-grams, which are N-sized letter collocations, are a well-used method in analysing written text (Cavnar and Trenkle 1994) and have been used in language classification. This paper attempts to create a dialect classifier based on character N-gram features. The chosen linguistic context is Middle English. The classifier successfully delineates between a cluster of Northern regions and a cluster of Midlands and Southern areas. Using this functioning text classifier, Middle English texts can be positioned in the regions they are most similar to. The analyses showed that texts of known origin are placed in a correct dialect cluster with high accuracy. Furthermore, Older Scots texts can be classified in relation to these Middle English regional clusters. Through this a quantitative confirmation that Older Scots is closest to Northern Middle English was found.

Keywords: Middle English, dialect classification, n-grams, corpus linguistics, bottom-up analysis

Buckley, Kevin. 2024. 'Uncovering linguistic lineage through using a character N-gram based dialect classifier'. In Christine Elsweiler (ed.). *The languages of Scotland and Ulster in a global context, past and present. Selected papers from the 13th triennial Forum for Research on the Languages of Scotland and Ulster, Munich 2021*. Aberdeen: FRLSU, pp. 139–76. ISBN: 978-0-9566549-7-7.

1 Introduction

Recent decades have seen the rise of quantitative methods applied to historical languages that display great dialectal variation (Piotrowski 2012). The chosen linguistic context of this paper is Middle English (ME) as it is known for its display of dialectal variation in the written word (McIntosh, Samuels and Benskin 1986). A Natural Language Processing method of analysing texts, character N-grams, is used to abstract gross dialect regions in ME in a bottom-up fashion, across Early ME (EME) and Late ME (LME). Firstly, the paper verifies these gross dialect regions through building a classifier, charting the accuracy in correctly disclosing the dialect of these texts. Secondly, this paper uses these working classifiers to investigate the linguistic lineage of Older Scots in relation to LME dialects. Through this, the relationship between Scots and Northern ME is quantitatively confirmed.

Table 1: Abbreviations

Abbreviation	Meaning
EME	Early Middle English
LAEME	*Linguistic Atlas of Early Middle English*
LALME	*Linguistic Atlas of Late Mediaeval English*
LAOS	*Linguistic Atlas of Older Scots*
LME	Late Middle English
ME	Middle English
MEG-C	*Middle English Grammar Corpus*
OE	Old English

2 Dialect classification

2.1 Top-down methods

Traditionally, dialects have been analysed in a top-down manner, guided by a linguistic expert in the language under study. However, these methods require large amounts of pre-processing, such as lemmatisation (headword tagging) or grammatical parsing, all of which require a thorough semantic understanding of the language under study. When dealing with large text corpora comprising of millions of words, it becomes impractical to rely on top-down methods requiring such pre-processing. Especially in historical languages where there is large variation, not all variables are known for certain, be it obscure spelling variants or noun forms. This slows down pre-processing efforts. Accordingly, this paper tries to apply methods that circumvent the need for laborious expert-led pre-processing of a text.

Quantitative analyses of dialect data have often used top-down methods. The field of Lexicostatistics and Dialectometry have used expert-compiled word lists to measure the genetic relationship between languages, through analysis of shared vocabulary (Millar and Trask 2015: 350). These disciplines use so-called Swadesh lists, often comprising 100 to 200 words, to compare vocabulary and generate a distance numeric. Dialectometry has used this method successfully. Nerbonne et al. (1996) used word lists of common words to measure the distance between Dutch dialects. The clusters[1] of dialects produced corresponded well to the dialect regions outlined by qualitative research. However, Buckley and Vogel (2019: 260) found that in the *Parsed Linguistic Atlas of Early Middle English* (Truswell et al. 2018), even for the entire corpus it was difficult to construct a Swadesh list, with

[1] In Dialectometry, it is popular to use clustering algorithms to produce genetic family trees or dendrograms from measurements of inter-language/inter-dialect distance (McMahon 2010; McMahon and Maguire 2012).

some lexical items only having one token. If it was difficult to construct a Swadesh list across a large sample of ME texts, then it would be certainly difficult for subgroupings such as an English county or any bespoke grouping. This study accordingly used a method that eschews the need for expert pre-processing, such as lemmatisation or creating Swadesh lists.

2.2 Character N-grams

Bottom-up methods, in comparison, are where features of a text are allowed to percolate to the top and are not selected by an expert. Wolk and Szmrecsanyi (2016) propose the use of bottom-up techniques for dialectology. They compare the use of bottom-up selected features versus pre-specified features in uncovering geolinguistic variation in the context of Modern English. They find that bottom-up methods yielded comparable results to top-down methods. Bottom-up techniques, instead of using research knowledge, use some facet of the features under consideration to select them for examination.

Character N-grams are a frequently used method of examining text data (Cavnar and Trenkle 1994). They are N-sized collocations of letters. These letter collocations are in essence fingerprints of a language, an abstract pattern of a language's usage. Cavnar and Trenkle (1994) see 2–3 slice character N-grams as capturing the most frequent words of a language along with their most important prefixes and suffixes. Character N-grams are often used in classification tasks. One such instance is language classification where the language of a test text is disclosed, such as in the program TextCat (Hornik et al. 2013), which perform language classification with high accuracy.

The Vector Space Model (Sidorov et al. 2014) represents languages as vectors, comprising the values of features, such as N-grams. As Damashek (1995) outlines, a written language sample can be represented as a vector whose components are the relative frequencies of the constituent N-grams of

the sample. Features are compared pairwise; a feature of vector A is compared across to the same feature of vector B. For the Vector Space Model, Cosine Similarity (Salton 1989) is a common metric of similarity. Cosine similarity computes the cosine of the angle between two vectors. It is calculated as the normalised dot product of two normalised vectors, i.e., the sum of the products of two equal length non-zero vectors. The resultant value ranges between 0 and 1, with 0 indicating no similarity at all and 1 indicating total similarity. The cosine similarity between two vectors *a* and *b* is as follows (Sidorov et al. 2014):

$$cosine(a,b) = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}}$$

Buckley and Vogel (2019) used this methodology to calculate the similarity between samples of medieval languages. Firstly, the similarity between epochs of historical English was calculated through a comparison of their N-gram frequency profiles, allowing for a measure of the amount of change within English over time. They also applied this method between languages, comparing historical English over time to Old French and Medieval Latin. This allowed for a quantification of the influence of French and Latin after the Norman Conquest (see section 3.1). The study showed that these bottom-up profiles can abstract the relationship between two languages that is found by qualitative research.

As these methods work between different time periods in a language and between languages, they can logically be extended to study the dialects of a language. Dipper and Schrader (2008) measured the cosine similarity between profiles of character N-grams in historical German dialects. The resulting cluster was successful in reproducing the distinction expected between Middle and Upper German and disclosed a cluster of Bavarian texts that was not detected by methods using whole word features.

As character N-grams have been used in language classification (see section 2.2), they can be applied to dialect classification. Dialect classification allows for the disclosure of the dialect that an inputted text is written in. This may allow for the identification of the dialect of texts that have not previously been identified by qualitative research. Dialect classification procedures have been used in many linguistic contexts to date. Ciobanu and Dinu (2016) classified Romanian dialects using both a word list and character N-gram slices derived from these word lists. They achieved above chance accuracy in distinguishing between dialects. Malmasi and Zampieri (2017) employed character N-grams from slice sizes 1–6 and word unigrams to identify Swiss German dialects. They found that generally character N-grams outperformed word unigrams, with larger and larger slices obtaining higher accuracy. Dialect identification has been also successful in the rich dialect context of Arabic. Zaidan and Callison-Burch (2014) used word unigrams and character N-grams to attempt to classify Arabic dialects. They found for their task, that word unigrams worked best, with 5-gram character slices the second most successful.

3 The linguistic context – Middle English

3.1 Periodisation and dialectal variation

The linguistic context chosen for this study is Medieval English, focusing on ME. ME is the descendant of Old English (OE), a Germanic language spoken in England up to 1066 (Mitchell and Robinson 2012). After 1066, the Norman Conquest induced profound effects on English (Freeborn 2006), with the ME period running up to 1500 (Brinton and Arnovick 2006). Within ME, two further subdivisions are made. Early ME (EME) is delineated as running from 1100–1340 and followed by Late ME (LME), comprising of the period from 1340–1500. Table 2 outlines the periodisation of medieval English.

ME is known for its dialectal variation. McIntosh et al. (1986) note that, by definition, all ME material before 1430 is considered dialectal. ME

has been claimed to have five main dialect regions, Northern, West Midlands, East Midlands, Southwestern, and Kentish (Brinton and Arnovick 2006: 242). Southwestern and Kentish can be grouped as Southern. Kentish can be referred to as Southeastern (Fulk 2012: 127). Northern ME was spoken north of the Humber River. The West and East Midlands dialects were spoken between the Thames and the Humber, the dividing line between them is a line north of Oxford. Kentish was spoken in Kent and part of Sussex. Southwestern is the remaining southern portion south of the Thames (Brinton and Arnovick 2006: 242).

Table 2: Periods of Historical English (adapted from Horobin and Smith, 2002: 1)

Period Name	**Time**
Old English	Up to 1100
Middle English	1100–1500
Early Middle English	1100–1340
Late Middle English	1240–1500
Early Modern English	From 1500

Access to ME is more indirect than Modern English with the study confined to written records (Milroy 1992: 161–162). There can be much dispute as to the pronunciation behind ME spelling, such as <a> giving no indication if it represents a low vowel, or front, or indeed its length or the degree of rounding (Milroy 1992: 163). For the current study, the focus is on ME orthographic variation. With the focus on bottom-up techniques with minimal top-down pre-processing, the focus is squarely on the data as it presents itself, e.g., the written word. No attempt to transliterate into a phonetic script was made. Buckley and Vogel when comparing the N-grams between OE and the related language Old Frisian found that orthography obscured the relationship, which became visible once the languages were transliterated into a phonetic script (2019: 289–293). This was not done for ME as it would introduce artificiality

into the results. Measured differences between regions would be due to the top-down insertions. As ME's variation is reflected in writing such as Northern <stan> vs. <stone> (see section 3.2), written text itself will be the focus.

3.2 Middle English features

This study, using character N-grams, takes a bird's eye view and is not focusing on any one feature or groups of features. The N-gram profile is an abstract fingerprint reflecting the sum usage of all words. It is difficult to examine features in isolation as multiple lexical items across several grammatical categories can contribute to one character N-gram feature. This fingerprint would feature contributions from all variation in a text, known and unknown, whether it be dialectal variation, genre, scribal idiolect, all impacting the character N-gram profile at once. However, given the use here of a frequency profile, the most dominant patterns should percolate to the surface and contribute more to the resulting similarity score. Here is a brief summation of ME's top dialectal features, firstly in written displays of phonological variation and secondly in the affixes of the dominant inflectional paradigms in ME.

EME long /a:/ is rounded to /ɔ:/ by 1225 (Fulk 2012). This change does not take place in Northern ME, which displays <a> with variants <ai> appearing. The rounded long vowel/ɔ:/ appears as <o>in south and midlands texts, hence Northern <ham> or <haim> compared to South <home>, and Northern <stan> compared to South <ston> 'stone' (Milroy 1992: 174; Fulk 2012).

OE <æ> becomes <a> except for Southeastern and West Midlands dialects, which often display <e>. OE <a> before a nasal, appears as <o> in West Midlands dialects (Milroy 1992: 175). OE <y> representing /y/ and /ʏ/, is spelled <u, ui, uy> in Southwestern and West Midlands ME (Brinton and Arnovick 2006; Fulk 2012), whereas in the south-east it appears as <e>, and

<i> elsewhere. Hence <brugge>, <bregge>, and <brigge> for OE *brycg* 'bridge' can be found.

Among consonants, Southern dialects and the southwest Midlands have word-initial <z> and <v> for OE /s/ and /f/ respectively, hence <zea> 'sea' and <vox> 'fox' (Milroy 1992: 175). Northern ME favours <g> or <k> where other dialects favour <ch>, hence <kirke> and <rigg> for <chirche> 'church' and <rigge> 'back'. OE <hw> has various outcomes in ME, some being <wh> and others being variants with <q>, such as <quh> (Milroy 1992: 175). <quh> and <qwh> are regular Northern spellings, hence Northern <quhilk> compared to London <which> (Kniezsa 1997: 32).

The dominant noun paradigms will undoubtedly feature in the character N-gram profile. ME had three types of noun paradigms that each features more in specific dialects. Table 3 lays out the suffixes from each type. From the start of ME, Northern dialects used only Type I. Midlands dialects used both Type I and II, and Southern dialects used all three (Mossé 1952: §55). ME is known for its vowel reduction (Brinton and Arnovick 2006: 266) and accordingly has only four unique N-grams for nouns, <e>, <s>, <es>, and <en>. Discrimination between dialects based on these suffixes may be unlikely.

Table 3: ME inflectional paradigms (adapted from Mossé 1952: §55; Fulk 2012: 58)

	Type I		**Type II**		**Type III**	
Singular		PDE *stone*		PDE *soul*		PDE *name*
Nom.	∅	*stōn*	e	*soule*	e	*name*
Acc.	∅	*stōn*	e	*soule*	e	*name*
Gen.	(e)s	*stōnes*	es	*soules*	e	*name*
Dat.	e	*stōn(e)*	e	*soule*	e	*name*
Plural						
all cases	(e)s	*stōnes*	es	*soules*	en	*namen*

The ME verbal paradigm offers more fruit at first glance, but on closer inspection there is a similar lack of unique features with many letter collocations reused (<eþ> appears as a plural marker for Southern dialects but as a 3rd person singular marker for the Midlands varieties). Table 4 shows the verbal inflectional paradigm per dialect.

Table 4: ME present tense indicative verb inflections, (adapted from Fulk 2012: 72)

	Southern & Kent	**West Midlands**	**East Midlands**	**Northern**
Singular				
1st	e	e	e	∅/e
2nd	(e)st	es(t)	est	es
3rd	(e)þ	eþ/es	eþ/es	es
Plural				
All persons	eþ	e(n)/es	e(n)/es	es/en

Another verbal variation that may prove more fruitful in discriminating dialects is the present participle suffix. The ending is generally <ande> or <and>, also in the North (Fernandez-Cuesta and Rodriguez-Ledesma 2006), but Southern texts display <inge> and <ynge> (Milroy 1992: 176; Ogura and Wang 2004).

All together this presents an uncertain picture for the efficacy of the character N-gram profile as a measure of variation. These features above appear to be contextual, in that the differences only appear when word tokens are aligned by their lemma, as in comparing Northern <stan> to <ston>, aligned by lemma ‘stone’. Character collocations may not capture this because the frequency of these letters may not be different as they appear in many other words. The character N-gram analysis used here must rely more on self-standing features that can stand out in a text without words being aligned by meaning. In this case it may be suffixes, such as on verbs (see

Table 4) and progressive participle suffixes such as <ande>. This study will examine to what degree ME dialects can be delineated without top-down alignment, using these self-standing features[2].

3.3 ME dialectology

ME dialect variation has been extensively charted by the sister projects *A Linguistic Atlas of Late Mediaeval English* (*LALME*) (McIntosh et al. 1986), covering LME, and *A Linguistic Atlas of Early Middle English* (*LAEME*) (Laing and Lass 2007), covering EME. These atlases use a questionnaire of items, i.e., linguistic features, to localise a text to a particular region. This is similar to a Swadesh list but is not confined to lexical items but also word parts such as suffixes.

The texts are localised by the 'fit-technique' (Benskin 1991), originally a technique carried out by hand (Laing and Williamson 2004: 88). Texts of known geographical origin, so-called 'anchor texts', had linguistic profiles (LPs) created from the questionnaire of items. Maps for each item in the questionnaire were created displaying the distribution of forms in the 'anchor texts' across geography. A text of unknown origin is assessed for which items on the questionnaire appear. Unlike Dialectometry, however, there is no measure of string distance applied between the word tokens in the text, but the method of *LALME* relies on exact string matching. The maps for these items are overlaid and the areas where the word tokens in the text do not occur are eliminated. Eventually an area of overlapping forms is identified. Through this the texts surveyed in *LALME* were localised to geographical locations. These localised texts were then added to the framework for fitting more texts (Stenroos and Thengs 2012: section 2).

[2] This study will use suffix N-grams (see section 4.3), hence word-initial variants such as <wh> will be less important.

There are several problems with this method. In order to cross compare multiple texts, the texts need to be placed in the same feature space (see section 2.2), such as a standardised list, e.g., a Swadesh list, or a vector of N-grams. The *LALME* method compares only the subset of items on the questionnaire that happen to appear. Thus, it is *test text > anchor text LPs* but this does not allow for cross-comparison with multiple other texts. They would need to be assessed on the same items. Also, the use of exact string matching over string distance, as in Dialectometry, does not allow the assessment of varying forms. Only shared word tokens between texts can be assessed and if the anchor texts lack these tokens, then informative dialectal forms in test texts will be ignored.

Furthermore, the adding of texts to the framework for further fitting will introduce cascading error. Wrongly localised texts will lead to further wrong localisation, which will then be also added to the framework with wrong localisation. This method is lacking the crucial step of verifying the localisation method by testing it on texts of known origin, in essence a dialect classification procedure. The situation is analogous to that of Old French. Dees (1980) produced an atlas based on ~3,300 charters that were what he calls primary witness documents (Dees 1988), texts that are localised by the text content. He assessed these on over 500 linguistic features (top-down collated). Dees (1987) localised Old French literary texts by comparison to these features. Interestingly, Dees provided a *Dees coefficient* that charts the degree to which a text conforms to the profile of a French region. De Jong (1996) used this coefficient to measure the degree to which Anglo-Norman charters diverged away from 13th century Anglo-Norman norms. A measurement such as this is lacking in *LALME*.

Both *LALME* and the Dees atlases lack verification of their method by assessing whether texts of known origin can be correctly placed in a region of origin. Dees (1992: 24) did attempt an early test on charters of known origin and found the results satisfactory but this was not reported in detail and

is not comprehensively assessing all regions in his atlases. Even modern efforts at assessing ME dialects do not verify their resulting clusters. Mäkinen (2019) tests using character 3-grams to examine ME texts. He did manage to achieve some clusters that abstracted some English counties. He found Cambridgeshire texts forming a cluster and Norfolk texts clusters. He struggled to differentiate dialects and genre with this method. This is probably due to assessing files individually. This current study will assess English counties as a whole, which will hopefully drown out less frequent features, with dialectal features common to all texts percolating to the top. Mäkinen (2020: 7) further went on to successfully abstract geographical distribution of counties, notably finding northern counties clustering together. However, as with *LALME*, these clusters are not verified by assessing the accuracy of placing texts of known origin to these clusters.

3.4. Older Scots

Historical or Older Scots is a descendent of Northern ME (Jones 1997). Before the 15th century, Northern ME and Older Scots were considered a common speech area (Williamson 2002). Furthermore, starting in the 16th century, Scots underwent a period of Anglicisation whereby Scots grew closer to Southern (Early Modern) English (Aitken 1985; Meurman-Solin 1997; Millar 2020: 86–91). The abandonment of Scots orthography is taken as sign of Anglicisation (Kniezsa 1997: 46). Distinctive Scots orthography was to a large degree replaced by an English orthography by the end of the 17th century in more official writing (Kniezsa 1997: 44).

Some of the features of Older Scots are the following: the trigraph <quh>, featuring in relative and interrogative pronouns such as <quhilk> ‘which’, <quhere> ‘where’, and <quhat> ‘what’ (Kniezsa 1997: 32; Hoffman 2019: 40; van Eyndhoven and Clark 2020), the present participle suffix <and>. Both of these two are Northern ME features (see section 3.2). Other Scots features are the plural suffix <is> and past tense verbal suffix <it>.

(Hoffman 2019: 40). They are replaced by <wh>, <ing>, <es> and <ed>, the ME Southern equivalents (see section 3.2). Hoffman reports an increase in these Anglicised equivalents after 1660 in examining the *Dunfermline Corpus* of Scots texts (Hoffman 2019: 52). Meurman-Solin (1997: 7–8) charts the frequency of Scots variants in the *Helsinki Corpus of Older Scots* (Meurman-Solin 1995), finding a drop in <quh> forms post 1600. Use of <it> over English <ed> on past tense of verbs is lower post 1600. Variants of 'they', <tha>, <thai>, and <thay>, begin to lower in usage after 1540.

3.5 Uses of a ME dialect classifier

With a functional ME dialect classifier, the region of ME texts found by top-down methods (*LALME* or Dialectometry) can be quantitatively confirmed or new knowledge can be created if the text has never been conclusively localised. Beyond this practical purpose, the analysis of related varieties of English can be carried out to see which areas of ME they are closest to and from that guess from which areas they originated. In this study, this will be termed 'lineage[3] detection'.

Dunning (1994) performed early experiments in using character N-gram profiles to identify the language of an inputted text. He noted that when his classification task was trained for the profile of English, inputted samples of German were classified as English over other languages in the training corpus. This indicated that German had more overlap in N-gram features with English than the other languages in the training corpus, due to its genetic inheritance with English. It was through this procedure that indications of the similarity between English and German could be glimpsed. Through this,

[3] Lexicostatistics (Millar and Trask 2015) seeks to abstract genetic relationships. This study does not claim its method can prove genetic relation, but it can show through shared N-gram features that ME texts are related. So, for descriptive purposes the term 'lineage' is being used.

texts of unknown or a suspected origin could be narrowed down to a linguistic neighbour through having the highest number of shared features. As this paper is looking solely at orthography, this would be a measure of shared graphies, with no comment on any underlying phonological overlap, as orthography can obscure phonological similarity (see section 3.1; Buckley and Vogel 2019).

Section 3.3 outlined the connection between Older Scots and Northern ME and also the Anglicisation of Scots. Given a functioning LME classifier, samples of Scots could be compared and their similarity to Northern LME disclosed. Furthermore, comparing later samples of Scots to LME dialects could chart the decrease in similarity to Northern ME and thereby chart the Anglicisation of Scots.

A previous study using a top-down Dialectometry method has shown that detecting the relation between Scots and ME is possible. McMahon and Maguire (2011) derived dialect distances between OE and ME dialects (as well as Modern English dialects), using a Swadesh list of common lexical items, transcribed phonetically. They found that the historical samples clustered together. They found two sub-branches of ME, a Southern branch and a branch containing East Midlands and Northern dialects. It was further found that Scots clustered with the historical English cluster, but it did not cluster closer to any sub-branch such as Northern ME. So, this lexicostatistic method did detect the relationship to historical English but didn't detect Northern ME as the closest relative.

4 Methodology and data

4.1 Aims of the study

The following are the main aims of the study:

1. Produce bottom-up N-gram-based clusters of ME counties to establish larger dialect groupings. These should delineate between some of the five ME dialects outlined in the literature.

2. Develop a character N-gram based classifier for EME and LME that can place texts in their correct dialect grouping with above chance accuracy.
3. Disclose the lineage between Older Scots and Northern ME and chart the divergence away from Northern ME over time, as per the Anglicisation of Scots from the 16th century (see section 3.4).

4.2 Corpora

Table 5 displays the corpora used in this study and the language or dialect used from the corpus. The ME materials were drawn from two sources. Firstly, LME dialectal material was drawn from the *Middle English Grammar Corpus* (*MEG-C*) (Stenroos et al. 2011). The texts of *LALME* were not transcribed previously. The *MEG-C* is a corpus comprising some of the ME texts that were surveyed in *LALME*. The corpus contained ~410 text files spanning from the early 1300s to a little after 1500. Text types range from legal or administrative documents to personal letters and samples of literary texts. As discussed in section 3.3, the localisations in the *MEG-C* descend from *LALME*. These are not primary witness documents[4] and their localisation to English counties by *LALME* leads to a house-of-cards effect. The resulting clusters below will be biased by any errors in the localisations by *LALME*.

Samples of EME dialect material were drawn from the online *LAEME* corpus collection (Laing and Lass 2007). The corpus contained 121 text files from the period 1150–1350. Samples of the LME London dialect were drawn

[4] The *Corpus of Middle English Local Documents* (*MELD*) (Stenroos et al. 2017) is a collection of extralinguistically localised texts but they are from 1400 onwards and this study aims to look at all of the ME period.

from the *Innsbruck Corpus of Middle English Prose* (Markus 2008). The corpus contained 32 texts that were annotated as the London dialect[5].

Samples of Early Modern English were taken from the *Helsinki Corpus of English Texts*. There were two sources of Older Scots. First, 80 texts were drawn from the *Helsinki Corpus of Older Scots* (Meurman-Solin 1995) and second, 1088 texts from the *Linguistic Atlas of Older Scots* (*LAOS*, Williamson 2013). The *Helsinki Corpus of Older Scots* inherits its periodisation from its mother corpus, the *Helsinki Corpus of English Texts* (Rissanen et al. 1991), but alternative periodisation has been proposed (Kopaczyk 2013: 252). However, for this initial exploratory study this periodisation will be kept.

Table 5: Corpora

Language or Dialect	Corpus	Source
EME	*Linguistic Atlas of Early Middle English*	Laing and Lass (2007)
LME	*Middle English Grammar Corpus*	Stenroos et al. (2011)
LME London	*Innsbruck Corpus of Middle English Prose*	Markus (2008)
EModE	*Helsinki Corpus of English Texts*	Rissanen et al. (1991)
Older Scots	*Helsinki Corpus of Older Scots*	Meurman-Solin (1995)
	Linguistic Atlas of Older Scots	Williamson (2013)

4.3 N-gram processing

It was decided to use word-final (suffix) slices of character N-grams. As mentioned in section 3.2, the N-gram profiles will heavily feature the

[5] It was indicated in the metadata that these files were not surveyed in *LALME* so there should be no overlap with the *MEG-C*.

inflectional morphology of a language and much of ME's variation is in its affixes, such as on nouns and verbs. Suffix N-grams were chosen in order to capture the fingerprint of these affixes (outlined in section 3.2). It is beyond the scope of this paper to test out the efficacy of other types (prefix, word N-grams, skip-grams). The efficacy of suffix N-grams was verified by top-down assessment, assessing how well the features abstracted the geographical distribution of ME regions (see Figure 1). As the cluster grouped geographical adjacent regions together, suffix n-grams were deemed satisfactory features in abstracting ME variation across space.

Suffix character N-grams were generated using the package Tau (Buchta et al. 2017). Profiles of suffixes were filtered, with features below a certain frequency of occurrence cut off. As per the Vector Space Model, N-gram feature frequency counts were assembled in a vector. Cosine similarity between vectors was calculated with the R package LSA (Wild 2009). Hierarchical clustering was performed in some analyses below using Pvclust (Suzuki and Shimodaira 2006). As clustering can be variable and sometimes unreplicable, Pvclust allows for the measurement of uncertainty in hierarchical clustering through repeated replication of the dendrogram cluster, giving the researcher insight into the reliability of the outputted cluster. Hierarchical clustering was performed using matrices of cosine similarity between the text samples.[6]

5 Dialect classification

5.1 Cluster of LME counties

Suffix N-gram profiles of LME counties were generated for each county in the *MEG-C*. A hierarchical cluster was performed on the resulting similarity matrices of county-to-county cosine similarity. Slice sizes 1–3 were used. A

[6] Each cluster presented below was bootstrap replicated 10,000 times.

similarity matrix for each slice size was calculated and an average of the three matrices was used for the cluster. The cluster revealed two branches of counties. Within each of these groupings, the procedure was repeated, and each cluster showed two branches. In total there were two gross county groupings with two subdivisions each, giving four regions in total. Table 6 lists the counties grouped into each region. Figure 1 graphically displays the geographical scope of each region.[7]

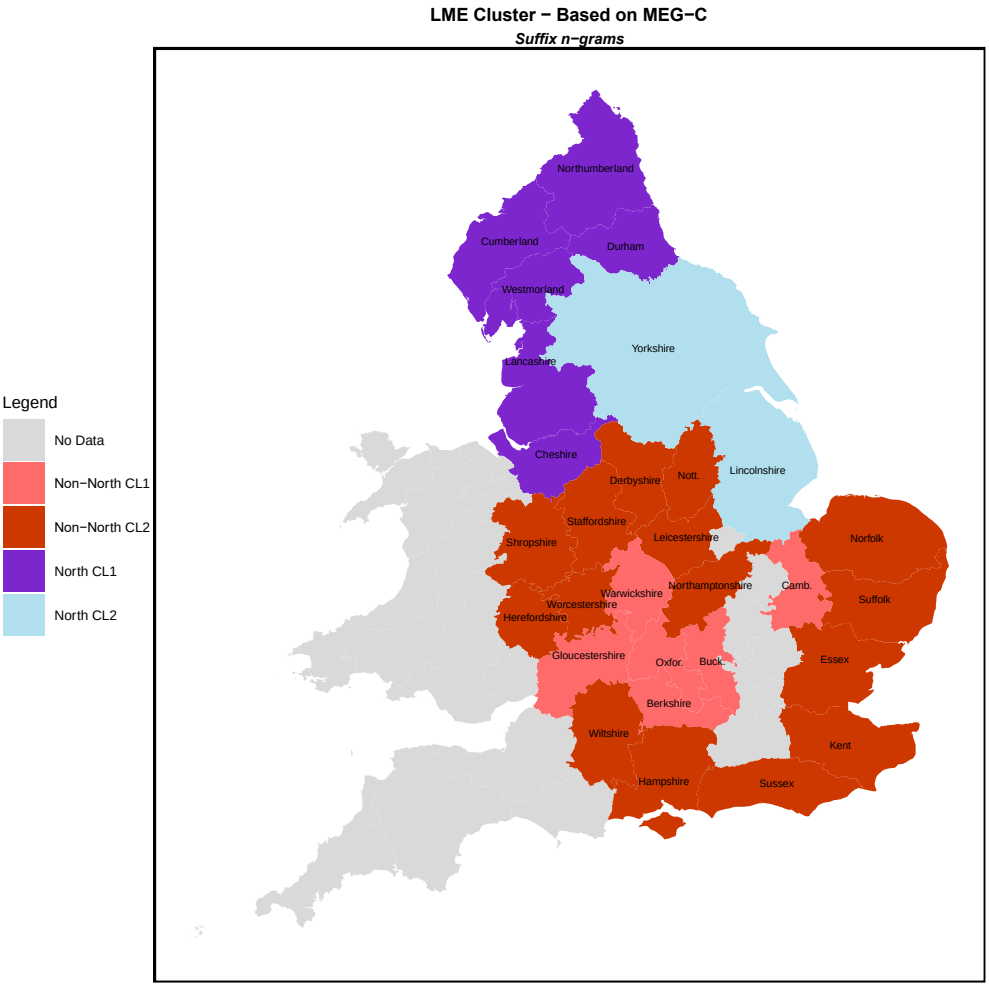

Figure 1: LME county cluster

[7] County boundaries taken from: https://www.county-borders.co.uk/.

Table 6: LME dialect clusters

County Name	Region	Sub Region	County Name	Region	Sub Region
Cheshire	North	CL1	Berkshire	Non-North	CL1
Cumberland	North	CL1	Buckinghamshire	Non-North	CL1
Durham	North	CL1	Cambridgeshire	Non-North	CL1
Lancashire	North	CL1	Gloucestershire	Non-North	CL1
Northumberland	North	CL1	Oxfordshire	Non-North	CL1
Westmoreland	North	CL1	Warwickshire	Non-North	CL1
Lincolnshire	North	CL2	Derbyshire	Non-North	CL2
Northern Yorkshire East Riding	North	CL2	Essex	Non-North	CL2
Yorkshire North Riding	North	CL2	Hampshire	Non-North	CL2
Yorkshire West Riding	North	CL2	Herefordshire	Non-North	CL2
York City Yorkshire	North	CL2	Kent	Non-North	CL2
North-West	North	CL2	Leicestershire	Non-North	CL2
			Norfolk	Non-North	CL2
			Northamptonshire	Non-North	CL2
			Nottinghamshire	Non-North	CL2
			Staffordshire	Non-North	CL2
			Sussex	Non-North	CL2
			Wiltshire	Non-North	CL2
			Worcestershire	Non-North	CL2
			Suffolk	Non-North	CL2
			Shropshire	Non-North	CL2

As can be seen, the first cluster abstracted a North/South divide with the northernmost counties forming one cluster and the regions of the midlands and south forming a second. This shows that regions where Northern ME was spoken were delineated from the rest of England. This cluster will be dubbed

'North' for this study. The south cluster contains regions from all other dialects, so will be named conservatively 'Non-North'.

The second round of clustering within each of these regions provides further divisions. Within 'North', there appears to be a west versus east divide, but these will be conservatively labelled North CL1 and North CL2. Within 'Non-North', there is a circle of counties surrounding a central region. These two subclusters both transect the East and West Midlands, and the region of Southern ME clusters also with Midlands counties. So this clustering fails to delineate succinctly between the 3 remaining qualitatively defined dialects. These two subdivisions will be dubbed Non-North CL1 and Non-North CL2.

5.2 EME counties cluster

Using *LAEME's* text files, the same procedure was carried out for the counties of EME. Table 7 outlines the counties in each region. Figure 2 shows the county groupings found.

Table 7: EME dialect clusters

County Name	Region	County Name	Region
Cheshire	CL1	Berkshire	CL2
		Cambridgeshire-Huntingdonshire	CL2
Essex	CL1		
Herefordshire	CL1	Gloucestershire	CL2
Shropshire	CL1	Kent	CL2
Worcestershire	CL1	Norfolk	CL2
		Somerset	CL2
		Wiltshire	CL2
		Yorkshire Ridings	CL2

These groupings map to the geography of England less well. However, some groupings can be discerned. There is a small western cluster with an outlier

county in the east, Essex. The second cluster contains all other counties. As no simplified labelling can be given, these clusters will be called EME CL1 and CL2.

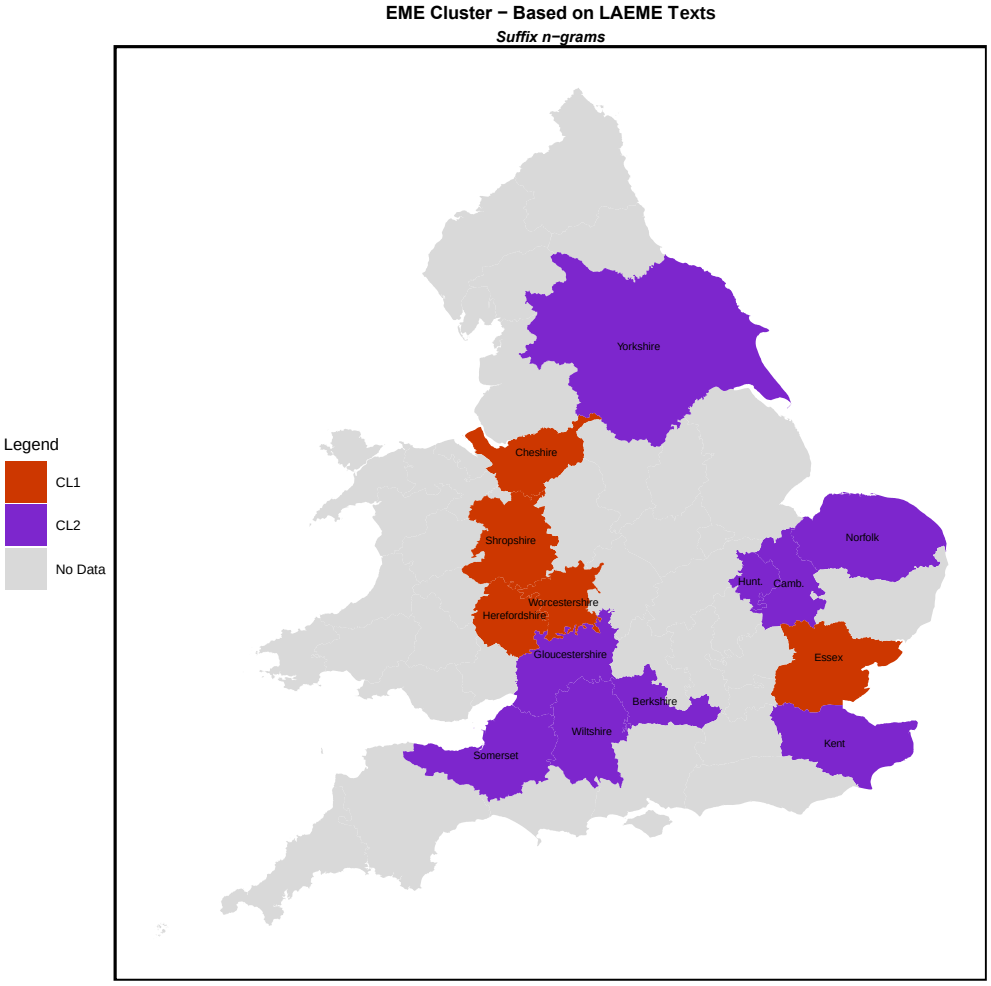

Figure 2: EME county cluster

5.3 Classifier procedure

The classifier procedure is as follows: the corpus' files are randomly divided into a training and test corpus at a 70/30 split.[8] Profiles of suffix N-grams

[8] This procedure is iterated 25 times so all data presented below is an average of all iterations.

(N=1–4) are generated for the dialect clusters. The suffix N-gram profile is compared through cosine similarity to the dialect region training profiles. The test file is labelled as the dialect cluster with which it has the highest cosine similarity.

5.3.1 LME dialect classifier accuracy – 'North' v. 'Non-North'

The *MEG-C* was used for LME. The 'North' and 'Non-North' dialect regions were used for this classifier (see Figure 1 and Table 6). Table 8 shows the accuracy per N-gram slice size of this classifier. All N-gram slice sizes were able to discriminate between the two clusters with suffix 4-grams being the most accurate.

Table 8: 'North' v. 'Non-North' classifier accuracy

	North Accuracy	**Non-North Accuracy**	**Balanced Accuracy**	**F1**
1-gram Suffix	0.65	0.73	0.69	0.64
2-gram Suffix	0.82	0.81	0.82	0.79
3-gram Suffix	0.77	0.92	0.84	0.81
4-gram Suffix	0.86	0.92	0.89	0.87

5.3.2 LME dialect classifier accuracy – 'North' CL1 v CL2

With the success in discriminating 'North' from 'Non-North' files, the subclusters of the 'North' dialect regions were used in a classifier. Table 9 shows the accuracy per N-gram slice size. Similarly, 4-grams are the most accurate in discriminating CL1 from CL2. However, the accuracy is lower than for 'North' v. 'Non-North', indicating there's less discriminability within the subclusters of North.

Table 9: LME 'North' Internal – CL1 v. CL2 classifier accuracy

	CL1 North Accuracy	**CL2 North Accuracy**	**Balanced Accuracy**	**F1**
1-gram Suffix	0.50	0.77	0.63	0.57
2-gram Suffix	0.56	0.76	0.66	0.62
3-gram Suffix	0.64	0.76	0.70	0.67
4-gram Suffix	0.69	0.82	0.76	0.73

5.3.3 LME dialect classifier accuracy – 'Non-North' CL1 v CL2

As with the 'North', the subclusters of the 'Non-North' region were used in a classifier. Table 10 shows the accuracy of the classifier per N-gram slice size. The classifier is accurate in discriminating Non-North CL1 from CL2 with 2-grams and 3-grams being the most accurate for these regions. This classifier displays a similar level of accuracy to the North Internal classifier, with 76 per cent and 78 per cent accuracy, respectively. There is a >10 per cent drop in discriminability for the dialect subclusters from the 89 per cent accuracy in discriminating the larger dialect regions, North v. Non-North.

Table 10: LME 'Non-North' Internal – CL1 v. CL2 classifier accuracy

	CL1 Accuracy	**CL2 Accuracy**	**Balanced Accuracy**	**F1**
1-gram Suffix	0.68	0.73	0.71	0.54
2-gram Suffix	0.82	0.74	0.78	0.63
3-gram Suffix	0.78	0.78	0.78	0.63
4-gram Suffix	0.76	0.76	0.76	0.61

5.3.4 EME dialect classifier accuracy

As with LME, an EME classifier was created using the clusters CL1 and CL2 (see Figure 2 and Table 7). Table 11 shows the accuracy in discriminating the clusters per N-gram slice size. The classifier is able to discriminate between

the clusters with a 4-gram slice being the most accurate. However, the 77 per cent accuracy is much lower than the 89 per cent accuracy found for LME. Nonetheless, there are two functioning classifiers than can place texts from EME and LME in large dialect clusters.

Table 11: EME – CL1 v. CL2 classifier accuracy

	CL1 Accuracy	**CL2 Accuracy**	**Balanced Accuracy**	**F1**
1-gram Suffix	0.76	0.62	0.69	0.70
2-gram Suffix	0.84	0.68	0.76	0.77
3-gram Suffix	0.82	0.67	0.74	0.76
4-gram Suffix	0.80	0.74	0.77	0.77

5.4 Verification test – *Helsinki Corpus* & *Innsbruck Corpus*

In order to further verify the accuracy of the classifiers, files of ME dialects from other sources are assessed by the classifier. The accuracy achieved by the classifiers could be a result of features unique to the corpus. To be of use, these classifiers must be able to place any ME text in its correct region.

ME dialect files from the *Helsinki Corpus* (Kytö 1996) and the *Innsbruck Corpus* (Markus 2008) were used. *Helsinki* files from the LME periods contained dialect labelling indicating whether a text was Northern ME or East/West Midlands or South. Northern ME files should be labelled as LME 'North' while the other files should be labelled as LME 'Non-North'. Table 12 displays the accuracy in correctly placing the dialect labelled *Helsinki* files into the correct LME dialect clusters (see section 4.2.1). The majority of *Helsinki* Northern files are placed correctly into the LME North cluster. Similarly, the majority of files from the other 3 dialects are placed in the Non-North cluster.

Table 12: *Helsinki Corpus* dialect files – 'North' v. 'Non-North' classifier

***Helsinki* Label**	**No. Files**	**Labelled North**	**Labelled Non-North**
Northern	4	0.75	0.25
EML, WML, Southern	31	0.26	0.74

The *Innsbruck Corpus* contains a large collection of files labelled as London. Given the geographical location, these files should be placed in the 'Non-North' cluster (see Figure 1). Table 13 displays in accuracy. All files are correctly placed in the 'Non-North' cluster, confirming the ability of the classifier to disclose the region of a ME text.

Table 13: *Innsbruck Corpus* London files – North v. Non-North classifier

***Innsbruck* Label**	**No. Files**	**Labelled North**	**Labelled Non-North**
London	32	0	1.0

5.5 Lineage detection

5.5.1 Lineage detection – Older Scots

As laid out in section 3.5, the ability of the classifier to disclose the lineage between Northern ME and Older Scots will be assessed. Samples of Older Scots will be assessed by the LME 'North' v. 'Non-North' classifier (see section 5.3.1). Scots texts should be labelled as the 'North' cluster.

Firstly, Scots files that are contemporaneous with the LME period were assessed. These were taken from the *LAOS*. Table 14 shows the amount of *LAOS* Scots files labelled as 'North'. The majority of Scots files are correctly identified North, suggesting the classifier can disclose the expected lineage.

Table 14: *LAOS* Scots files – North v. Non-North classifier

***LAOS* period**	**No. Files**	**'Non-North'**	**'North'**
1350–1495	1088	0.16	0.84

With Scots materials contemporaneous with LME being disclosed as similar to North, the Anglicisation of Scots (Aitken 1985; Meurman-Solin 1997) away from Northern forms was assessed. Samples of Scots over time were taken from the *Helsinki Corpus of Older Scots* (Meurman-Solin 1995). As a control, samples of Early Modern English were taken from the *Helsinki Corpus of English Texts* (Kytö 1996). Table 15 shows the number of files labelled as LME North or Non-North.

Firstly, Early Modern English from just after the LME period (1500–1570) is correctly identified as majority LME 'Non-North'. Surprisingly, over time, there is a growth in the number of files being labelled as North, suggesting that some innovations in Early Modern English may overlap with features in Northern ME.[9] Nevalainen and Raumolin-Brunberg (2017: 177–180) examined the growth in LME Northern features in the south. They charted the growth in the use of the Northern 3rd person singular indicative suffix *-es* over Southern *-eth* as well as the use of the Northern possessive pronouns *my* and *thy* over Southern *mine* and *thine*. Each feature displays high usage in Northern texts at the end of the ME period (1460–1499). Southern texts show a steady increase in usage of these features during the 17th century. These LME Northern features in 17th century texts could explain the 19% of EModE texts labelled as Northern. Further research should precisely tag the number of Northern features in such texts to correlate their number with this measured classification.

[9] Early Modern English files carry no dialect labelling so the proportion of texts written in the north cannot be assessed.

For Scots, files contemporaneous with LME are found to be majority 'North' as with the *LAOS* files. However, the number of files labelled as North is lower, with 40 per cent of the texts labelled as 'Non-North'. Scots from just after the LME period (1500–1570) are majority labelled as 'North', with 88 per cent 'North'. The classifier picks up on the Anglicisation of Scots over time as the number of files labelled as 'North' decreases sharply over time. 70 per cent of 1640–1700's Scots are now found to be more similar to LME 'Non-North', indicating a loss of forms that originally overlapped with LME 'North' in the periods 1450–1500 and 1500–1570.

Table 15: Older Scots and Early Modern English – 'North' v. 'Non-North' classifier

Scots				***Early Mod. Eng.***			
***Helsinki* Epoch**	**No. Files**	**Non-North**	**North**	***Helsinki* Epoch**	**No. Files**	**Non-North**	**North**
1450–1500	10	0.4	0.6				
1500–1570	17	0.12	0.88	*1500–1570*	56	0.91	0.09
1570–1640	29	0.38	0.62	*1570–1640*	50	0.88	0.12
1640–1700	23	0.71	0.29	*1640–1710*	52	0.81	0.19

6 Discussion

As per the aims laid out in section 4.1, dialect clusters that abstract the delineations expected from the literature were created. The major distinction found was between LME northern counties and midlands/south counties. However, subclusters did not map to any of the further ME dialects, i.e., clusters did not delineate between East and West Midlands regions.

Classifiers were made that successfully discriminated between these clusters and subclusters with above-chance accuracy. The most accurate discrimination was between the large dialect clusters LME 'North' and 'Non-North' but discrimination between subclusters was still above 75 per cent.

The lineage between Older Scots and Northern ME was successfully found by comparing Scots texts to the LME North and Non-North clusters. Contemporaneous Scots and Scots from just after the LME period were majority identified as LME North, thereby disclosing and confirming the expected lineage.

All these provide a method to disclose the origin of ME texts of unknown or disputed origin. Also, it can provide a method of assessing the origin of other varieties of English. For instance, the medieval variety of English found in Ireland could be assessed in a similar manner to Scots. *The Kildare Poems* are a selection of medieval English poems written in Ireland (Hickey 2003). Comparison of these texts to the classifiers, discloses that they are most similar to LME 'Non-North' and within that they are most similar to 'Non-North' CL2. This shows that for these related varieties the region of linguistic origin can be assessed.

However, the clusters created here are still quite large, with identification down to a county level not attempted. Further research can examine at what point discrimination between smaller and smaller regions of medieval England breaks down when using profiles of character N-grams. Further studies should compare the ability to localise texts to a county level by word lists (such as *LALME* questionnaire lists) and by bottom-up lists of features, be it character N-grams or other features such as word unigrams.

Lastly, there is an implicit bias in the methodology used here. These classifiers operate on the principle that the more features of a dialect a text has, the more likely it is to be that dialect. However, this has no way to assess shibboleths. There can be single features that prove the dialect of a text, but these would be outweighed by other features. For instance, if a text displays <stan> over <ston>, this suggests a Northern text but if there are numerically more features that are non-Northern then the text will be wrongly classified. Classification procedures should include an assessment of unique dialect features that would outweigh other features.

In conclusion, the study has been broadly successful in creating classifiers to disclose the dialect of a ME text and the lineage between Older Scots and LME North has been confirmed.

References

Aitken, Adam. J. 1985. 'A history of Scots: *Concise Scots Dictionary*'. In Robinson (1985): ix–xiii.

Alcorn, Rhona, Joanna Kopaczyk, Bettelou Los, and Benjamin Molineaux (eds.) 2019. *Historical dialectology in the digital age*. Edinburgh University Press.

Benskin, Michael. 1991. 'The "fit"-technique explained.' In Riddy (1991): 9–26.

Bös, Birte, and Claudia Claridge (eds.) 2019. *Norms and conventions in the history of English.* Amsterdam: Benjamins.

Blake, Norman (ed.) 1992. *The Cambridge history of the English language. Volume II: 1066–1476*. Cambridge: Cambridge University Press.

Brinton, Laurel J., and Leslie K. Arnovick. 2006. *The English language: A linguistic history*. Oxford: Oxford University Press.

Buchta, Christian, Kurt Hornik, Ingo Feinerer, and David Meyer. 2017. *Package 'tau'*. N.p. https://cran.r-project.org/web/packages/tau/index.html [accessed: 18 July 2023].

Buckley, Kevin, and Carl Vogel. 2019. 'Using character n-grams to explore diachronic change in medieval English'. *Folia Linguistica* 53(s40–s2): 249–299.

Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.) 2016. *Proceedings of the Tenth International Conference on Language Resources and*

Evaluation (LREC 2016). Paris: European Language Resources Association (ELRA).

Cavnar, William B., and John M. Trenkle. 1994. 'N-gram-based text categorization'. In Cavnar and Trenkle (1994): 161–175.

Cavnar, William. B., and John. M. Trenkle. 1994. *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval.*

Ciobanu, Alina Maria, and Liviu P. Dinu. 2016. 'A Computational Perspective on the Romanian Dialects'. In Calzolari et al. (2016): 3281–3285.

Côté, Marie-Hélène, Remco Knooihuizen, and John Nerbonne. 2016. *The future of dialects: Selected papers from Methods in Dialectology XV*. Berlin: Language Science Press.

Cruickshank, Janet, and Robert McColl Millar (eds.) 2013. *After the storm: Papers from the Forum for Research on the Languages of Scotland and Ulster Triennial Meeting, Aberdeen 2012.* Aberdeen: Forum for Research on the Languages of Scotland and Ulster.

Damashek, Marc. 1995. 'Gauging similarity with n-grams: Language-independent categorization of text'. *Science* 267(5199): 843–848.

Dees, Anthonij. 1980. *Atlas des formes et des constructions des chartes françaises du 13e siècle.* Tübingen: Niemeyer.

Dees, Anthonij. 1987. *Atlas des formes linguistiques des textes littéraires de l'ancien français.* Max Niemeyer Verlag. Tübingen: Niemeyer.

Dees, Anthonij. 1988. 'Propositions for the study of Old French and its dialects'. In Fisiak (1988): 139–148.

Dees, Anthonij. 1992. 'Les chartes dans la recherche linguistique et philologique'. *Le médiéviste et l'ordinateur* 25(1): 23–27.

De Jong, Thera. 1996. 'Anglo-French in the 13th and 14th centuries: Continental or insular dialect?'. In Nielsen and Schøsler (1996): 55–70.

Denison, David, Ricardo Bermúdez-Otero, Chris McCully, and Emma Moore (eds.) 2012. *Analysing older English*. Cambridge: Cambridge University Press.

Dipper, Stefanie, and Bettina Schrader. 2008. 'Computing distance and relatedness of medieval text variants from German'. In Storrer et al. (2008): 39–51.

Dossena, Marina, and Roger Lass (eds.) 2004. *Methods and data in English historical dialectology*. Bern: Peter Lang.

Dossena, Marina, Richard Drury, and Maurizio Gotti (eds.) 2008. *English historical linguistics 2006, vol. 3: Geo-historical variation in English*. Amsterdam and Philadelphia: Benjamins.

Dunning, Ted. 1994. *Statistical identification of language*. Las Cruses, NM: Computing Research Laboratory, New Mexico State University.

Durieux, Gert, Walter Daelemans, and Steven Gillis (eds.) 1996. *CLIN VI: Papers from the Sixth CLIN Meeting (Antwerp, December 1, 1995)*. Antwerp: UIA Center for Dutch Language and Speech.

e-LALME = An Electronic Version of A Linguistic Atlas of Late Mediaeval English. 2013. Compiled by Michael Benskin, Margaret Laing, Vasilis Karaiskos, and Keith Williamson. Edinburgh: The University of Edinburgh. http://www.lel.ed.ac.uk/ihd/elalme/elalme.html [accessed: 19 January 2018].

Van Eyndhoven, Sarah, and Lynn Clark. 2020. 'The <quh->–<wh-> switch: An empirical account of the anglicisation of a Scots variant in Scotland during the sixteenth and seventeenth centuries'. *English Language & Linguistics* 24(1): 211–236.

Fanego, Teresa, Belén Méndez-Naya, and Elena Seoane (eds.) 2002. *Sounds, words, texts and change: Selected Papers from 11 ICEHL, Santiago de Compostela, 7–11 September 2000*. Amsterdam and Philadelphia: Benjamins.

Fernández Cuesta, Julia, and Maria Nieves Rodríguez Ledesma. 2008. 'Northern Middle English: Towards telling the full story'. In Dossena, Drury and Gotti (2008): 91–109.

Fisiak, Jacek (ed.) 1988. *Historical dialectology: Regional and social*. Berlin, New York, and Amsterdam: Mouton de Gruyter.

Fulk, Robert D. 2012. *An introduction to Middle English: Grammar and texts*. Peterborough, ON: Broadview Press.

Freeborn, Dennis. 2006. *From Old English to standard English*. 3rd ed. Basingstoke: Macmillan.

HC = The Helsinki Corpus of English Texts. 1991. Compiled by Matti Rissanen, Merja Kytö, Lena Kahlas-Tarkka, Matti Kilpiö, Saara Nevanlinna, Irma Taavitsainen, Terttu Nevalainen, and Helena Raumolin-Brunberg. Helsinki: Department of Modern Languages, University of Helsinki.

HCOS = The Helsinki Corpus of Older Scots. 1995. Compiled by Anneli Meurman-Solin. Helsinki: Department of Modern Languages, University of Helsinki.

Hickey, Raymond. 2003. 'A corpus of Irish English'. In Hickey (2003): 237–249.

Hickey, Raymond. 2003. *Corpus presenter: Software for language analysis – with a manual and a corpus of Irish English as sample data*. Amsterdam and Philadelphia: Benjamins.

Hickey, Raymond (ed.). 2010. *The handbook of language contact*. Malden, MA: Wiley-Blackwell.

Hoffman, Klaus. 2019. 'Approaching transition Scots from a micro-perspective: The Dunfermline corpus, 1573–1723'. In Alcorn et al. (2019): 39–60.

Hornik, Kurt, Patrick Mair, Johannes Rauch, Wilhelm Geiger, Christian Buchta, and Ingo Feinerer. 2013. 'The textcat package for n-gram based text categorization in R'. *Journal of Statistical Software* 52(6): 1–17.

Horobin, Simon, and Jeremy J. Smith. 2002. *An introduction to Middle English*. New York: Oxford University Press.

Jones, Charles (ed.) 1997. *The Edinburgh history of the Scots language*. Edinburgh: Edinburgh University Press.

Kay, Christian, and Jeremy J. Smith (eds.) 2004. *Categorization in the history of English*. Amsterdam and Philadelphia: Benjamins.

Kniezsa, Veronika. 1997. 'The origins of Scots orthography'. In Jones (1997): 24–46.

Kopaczyk, Joanna. 2013. 'Rethinking the traditional periodisation of the Scots language'. In Cruickshank and Millar (2013): 233–60.

Kytö, Merja. 1996. *Manual to the diachronic part of the Helsinki Corpus of English Texts*: *Coding conventions and lists of sources.* Helsinki: University of Helsinki, Department of English. http://korpus.uib.no/icame/manuals/HC/INDEX.HTM [accessed: 27 September 2017].

LAEME = A Linguistic Atlas of Early Middle English, 1150–1325. Version 3.2. 2007. Compiled by Margaret Laing and Roger Lass. Edinburgh: The University of Edinburgh. http://www.lel.ed.ac.uk/ihd/laeme2/laeme2_framesZ.html [accessed: 27 September 2017].

Laing, Margaret, and Keith Williamson. 2004. 'The archaeology of medieval texts'. In Kay and Smith (2004): 85–146.

LAOS = A Linguistic Atlas of Older Scots, Phase 1: 1380–1500. Version 1.2. 2013. Compiled by Keith Williamson. Edinburgh: The University of Edinburgh. http://www.lel.ed.ac.uk/ihd/laos1/laos1.html [accessed: 14 September 2019].

Malmasi, Shervin, and Marcos Zampieri. 2017. 'German dialect identification in interview transcriptions'. In Nakov et al. (2017): 164–169.

Markus, Manfred. 2008. *Innsbruck Corpus of Middle English Prose: Collection of Middle English texts compiled for ICAMET*. Innsbruck: University of Innsbruck.

Mäkinen, Martti. 2019. 'Testing a stylometric tool in the study of Middle English documentary texts'. In Bös and Claridge (2019): 149–166.

Mäkinen, Martti. 2020. 'Stylo visualisations of Middle English documents'. *Journal of Data Mining & Digital Humanities: Special issue on Visualisations in Historical Linguistics*: 1–10.

McIntosh, Angus, Michael L. Samuels, and Michael Benskin. 1986. 'General introduction, chapter 1: Orientation'. In McIntosh, Samuels and Benskin (1986): 155–116.

McIntosh, Angus, Michael L. Samuels, and Michael Benskin. 1986. *A Linguistic Atlas of Late Medieval English, vol. 1: General Introduction, Index of Sources, Dot Maps*. Aberdeen: Aberdeen University Press. http://www.lel.ed.ac.uk/ihd/elalme/intros/atlas_gen_intro.html [accessed: 19 January 2018].

McMahon, April. 2010. 'Computational models and language contact'. In Hickey (2010): 128–147.

McMahon, April, and Warren Maguire. 2012. 'Quantitative historical dialectology'. In Denison et al. (2012): 140–158.

MEG-C = The Middle English Grammar Corpus. Version 2011.1. 2011. Compiled by Merja Stenroos, Martti Mäkinen, Simon Horobin, and Jeremy Smith. Stavanger: University of Stavanger. https://www.uis.no/en/middle-english-grammar-corpus-meg-c-0 [accessed: 17 July 2023].

MELD = A Corpus of Middle English Local Documents (MELD), version 2017.1. 2017. Compiled by Merja Stenroos, Kjetil V. Thengs, and Geir Bergstrøm. Stavanger: University of Stavanger. https://www.uis.no/en/meld-corpus-files [accessed: 17 July 2023].

Meurman-Solin, Anneli. 1995. ‘A new tool: *The Helsinki Corpus of Older Scots (1450–1700)*’. *ICAME Journal* 19(4): 49–62.

Meurman-Solin, Anneli. 1997. ‘Differentiation and standardisation in early Scots’. In Jones (1997): 3–23.

Millar, Robert McColl, and Robert L. Trask (eds.) 2015. *Trask's historical linguistics*. 3rd ed. London and New York: Routledge.

Millar, Robert McColl. 2020. *A sociolinguistic history of Scotland*. Edinburgh: Edinburgh University Press.

Milroy, James. 1992. ‘Middle English dialectology’. In Blake (1992): 156–204.

Mitchell, Bruce, and Fred C. Robinson. 2012. *A guide to Old English*. 8th ed. Malden, MA: Wiley-Blackwell.

Mossé, Fernand. 1952. *A handbook of Middle English*. Translated by James A. Walker. Baltimore: The John Hopkins Press.

Nakov, Preslav, Marcos Zampieri, Nikola Ljubešić, Jörg Tiedemann, Shevin Malmasi, and Ahmed Ali (eds.) 2017. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Stroudsburg, PA: The Association for Computational Linguistics.

Nerbonne, John, Wilbert Heeringa, Erik van den Hout, Peter van der Kooi, Simone Otten, and Willem van de Vis. 1996. ‘Phonetic distance between Dutch dialects’. In Durieux, Daelemans and Gillis (1996): 185–202.

Nerbonne, John, and William A. Kretzschmar Jr. 2013. ‘Dialectometry++’. *Literary and Linguistic Computing* 28(1): 2–12.

Nevalainen, Terttu, and Helena Raumolin-Brunberg. 2017. *Historical sociolinguistics: Language change in Tudor and Stuart England*. London and New York: Routledge.

Nielsen, Hans Frede, and Lene Schøsler. 1996. *The Origins and development of emigrant languages: Proceedings from the Second Rasmus Rask*

Colloquium, Odense University, November 1994. Amsterdam and Philadelphia: Benjamins.

Ogura, Mieko, and William SY Wang. 2004. 'Dynamic dialectology and complex adaptive system'. In Dossena and Lass (2004): 1–34.

Piotrowski, Michael. 2012. *Natural language processing for historical texts*. San Rafael, CA: Morgan and Claypool.

PLAEME = A Parsed Linguistic Atlas of Early Middle English. 2018. Compiled by Robert Truswell, Rhona Alcorn, James Donaldson, and Joel Wallenberg. Edinburgh: The University of Edinburgh. https://datashare.is.ed.ac.uk/handle/10283/3032 [accessed: 17 April 2018].

Riddy, Felicity (ed.) 1991. *Regionalism in late medieval manuscripts and texts: Essays celebrating the publication of* A Linguistic Atlas of Late Mediaeval English. Cambridge and Rochester, NY: D.S. Brewer.

Robinson, Mairi (ed.). 1985. *Concise Scots Dictionary*. Edinburgh: Edinburgh University Press.

Salton, Gerard. 1989. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Sidorov, Grigori, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. 'Soft similarity and soft cosine measure: Similarity of features in vector space model'. *Computación y Sistemas* 18(3): 491–504.

Stenroos, Merja, and Kjetil V. Thengs. 2012. 'Two Staffordshires: Real and linguistic space in the study of late Middle English dialects'. In Tyrkkö et al. (2012). https://varieng.helsinki.fi/series/volumes/10/stenroos_thengs/ [accessed: 17 July 2023].

Storrer, Angelika, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner (eds.) 2008. *Text resources and lexical knowledge: Selected*

papers from the 9th Conference on Natural Language Processing (KONVENS-08). Berlin: Mouton de Gruyter.

Suzuki, Ryota, and Hidetoshi Shimodaira. 2006. 'Pvclust: An R package for assessing the uncertainty in hierarchical clustering'. *Bioinformatics* 22(12): 1540–1542.

Tyrkkö, Jukka, Matti Kilpiö, Terttu Nevalainen, and Matti Rissanen (eds.) 2012. *Outposts of historical corpus linguistics: From the Helsinki Corpus to a proliferation of resources.* Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/10/index.html [accessed: 17 July 2023].

Wild, Fridolin. 2009. *The lsa package*. N.p. https://github.com/cran/lsa [accessed: 17 July 2023].

Williamson, Keith. 2002. 'The dialectology of "English" north of the Humber, c. 1380–1500'. In Fanego, Méndez-Naya and Seoane (2002): 253–286.

Wolk, Christoph, and Benedikt Szmrecsanyi. 2016. 'Top-down and bottom-up advances in corpus-based dialectometry'. In Côté, Knooihuizen and Nerbonne (2016): 225–243.

Zaidan, Omar F., and Chris Callison-Burch. 2014. 'Arabic dialect identification'. *Computational Linguistics* 40(1): 171–202.