Kevin Buckley* and Carl Vogel

Using character N-grams to explore diachronic change in medieval English

https://doi.org/10.1515/flih-2019-0012
Received 30 April 2018; revision invited August 23, 2018; revision received May 03, 2019; accepted June 07, 2019

Abstract: This paper applies character N-grams to the study of diachronic linguistic variation in a historical language. The period selected for this initial exploratory study is medieval English, a well-studied period of great linguistic variation and language contact, whereby the efficacy of computational techniques can be examined through comparison to the wealth of thorough scholarship on medieval linguistic variation. Frequency profiles of character N-gram features were generated for several epochs in the history of English and a measure of language distance was employed to quantify the similarity between English at different stages in its history. Through this a quantification of internal change in English was achieved. Furthermore similarity between English and other medieval languages across time was measured allowing for a measurement of the well-known period of contact between English and Anglo-Norman French. This methodology is compared to traditional lexicostatistical methods and shown to be able to derive the same patterns as those derived from expert-created feature lists (i.e. Swadesh lists). The use of character N-gram profiles proved to be a flexible and useful method to study diachronic variation, allowing for the highlighting of relevant features of change. This method may be a complement to traditional qualitative examinations.

Keywords: computational linguistics, history of English, language contact, diachronic linguistics, character N-grams

***Corresponding author: Kevin Buckley,** School of Modern Languages, Newcastle University, Newcastle upon Tyne, UK, E-mail: kebuckle@tcd.ie
Carl Vogel, Trinity Centre for Computing and Language Studies, Computational Linguistics Group, School of Computer Science and Statistics, Trinity College Dublin, the University of Dublin, Dublin, Ireland, E-mail: vogel@tcd.ie

1 Introduction

1.1 General introduction

The study of language variation and change has traditionally applied qualitative methods but recent decades have seen a rise in the application of quantitative methods to the study of linguistic phenomena, including historical (Piotrowski 2012). The rise of computer readable corpora for many historical and modern languages, often comprising millions of words, has opened the door to apply quantitative methods to the study of language variation. This recent trend of using ‘Big Data’ offers the exciting possibility of not only confirming the findings of the laborious and rigorous qualitative research of the previous centuries but of complementing these findings by discovering new patterns not easily discernible on a smaller scale.

This current paper attempts, by using a computational method known as character N-gram models, to study diachronic linguistic variation at a finer temporal resolution in the context of historical English. Through profiles of a language’s constituent character N-gram, an abstraction of a historical language was created at various points in its history and a measure of distance quantified the degree of change between periods. This allowed for the quantification of changes in orthographic and phonetic distance within the language over time but also between languages providing a quantification of language contact between English and French in the medieval period. The aim of this study is to examine the efficacy of this method to replicate the linguistic trends discovered through qualitative studies of English.

1.2 The linguistic context of the present study

Historical English was chosen as the period of interest in this study as it presents a well-known period of linguistic variation due to language contact between English and Norman French. The rigorous scholarship on this period of early English presents a benchmark against which the results of quantitative methods can be assessed. Bottom-up methods are where fewer assumptions and delineations or filters decided by the researcher are applied to the data, in effect letting the data speak for itself. These bottom-up methods can therefore be compared against the discoveries of traditional qualitative research to examine how sensitive or accurate a measure is in detecting well-known phenomena.

The relevant features of the two languages in contact will be briefly summarized. Old English (OE) was the vernacular Germanic language spoken in England from 449 to 1066 (Mitchell and Robinson 2011). After 1066, the Norman Conquest induced profound effects on English (Freeborn 2006). Middle English (ME) is defined as the period lying after the Norman Invasion, which brought a migration of Anglo-Norman speakers to England.

Burrow and Turville-Petre (2013) point out that the most distinguishing factor of ME is an increasingly mixed word stock derived from French. The English lexicon underwent drastic changes during ME with a large influx of loanwords from Anglo-Norman. Durkin (2014) attempts to quantify the contribution of loanwords from French with an analysis of the Middle English Dictionary, which provides coverage of the lexicon up to the fifteenth century. Among new additions in the ME period, he found that 48% were of Latin and French origin. Durkin's analysis demonstrates that French loanwords first show a peak in the early fourteenth century and continue to enter the lexicon through to the fifteenth century, with the influx reaching its highest peak between 1350–1450. Furthermore Durkin notes that French loanwords are scarce among the basic grammatical words of English, i.e. closed category words (Cinque 2006).

This import of French vocabulary did not just affect the lexicon, but produced many changes in orthography. French scribes imported letters such as <z> and <v> for /z/ and /v/ (Lass 2000), as well as vowel representations such as <ou> for OE <u> /u:/. Dalton-Puffer (1996) furthermore examines the effect on English's morphology. English imported several French suffixes and prefixes, demonstrating further change outside of the lexicon. See Table 1 for abbreviations used throughout.

Table 1: Abbreviations.

Abbreviation	Meaning
AN	Anglo-Norman
CF	Continental French
HC	Helsinki corpus of English texts
IDS	Intercontinental dictionary series
LAEME	Linguistic atlas of early middle English
ME	Middle English
OE	Old English
P-LAEME	The parsed linguistic atlas of early middle English

1.3 Language distance and Lexicostatistics

A core concept to this study is the concept of Language Distance or inter-language similarity. According to Borin (2013), the judgment of similarity between languages often focuses on “symbol sequences”, comparing the linguistic surface features such as orthography or a phonetic transcription. One of the first fields attempting to measure inter-language distance was Lexicostatistics, which attempts to measure the genetic relationship between languages through the statistical analysis of vocabularies (Millar and Trask 2015).

Lexicostatistical studies aim to provide a quantitative measurement of an abstract distance between a pair of languages, which is often conceptualized as a genetic distance. These studies employ word-sense lists or Swadesh lists comprising each language’s lexical items for a particular word-sense. Through shared cognates in each language a quantification of the languages shared inheritance can be achieved. A measure of distance is used to quantify the difference between each word pair. The measure traditionally used is the Levenshtein or edit-distance, or variations thereof, which measures the number of insertions or deletions between a word pair (Levenshtein 1966; Rama et al. 2015; Wichmann et al. 2010).

1.4 Problems in applying word-sense lists to studying historical variation

There are several issues in applying lexicostatistical methods to historical languages and for using them to examine internal change across time and for measuring contact-induced change. These fall into three main groupings, problems in the creation of the lists, the implicit or explicit purpose of the lists and the effect on their contents, and issues in applying word-to-word comparisons between the lists.

Rama et al. (2015) note that the contents of Swadesh lists are based on human expert cognacy judgments. These cognacy judgments require experts for each of the languages under consideration. The creation of a list can be a laborious effort, tracing the etymology of a word, and with modern languages consulting native speakers to assess the match of a lexical unit to a word-sense.

In the collation of lexical items, a bias can be inadvertently given to one dialect over others and one time epoch over another. The cognate lists being drawn from dictionaries will inherit the bias of the dictionary chosen. For instance, for Old English, the written record presents a dialectal bias towards West Saxon and a temporal bias to the period of greatest attestation. A word-sense list would be problematic to construct for languages with great variation. For Middle English, Horobin and Smith (2002) outline that, within a temporal epoch, several dozen word forms can be found for one word sense alone, for the word-sense "through", <throgh>, <throw> <þorow> or even spellings such as <yora>, <yruȝ> or <trghug> can be found. Swadesh lists typically chosen one variant. For the Old English word for the word-sense "king", is <cyng> or <cyning> chosen? For interchangeable spelling variants, <þæt> and <ðæt>, which is favoured? The Intercontinental Dictionary Series (IDS; Key and Comrie 2015; Borin et al. 2013) can give multiple entries for one word sense. However no frequency of usage information is present in helping one decide which form is the most appropriate to select for the word sense. All entries for a word sense could be considered and their scores averaged, yet this would privilege less frequent words, putting their effect on the resulting distance on equal footing with the most used word. If frequency information was available a weighting to each word according to their frequency could be applied, thereby lowering the effect of less frequent items on the distance numeric.

A second problem is that lexicostatistics has the explicit goal of uncovering a genetic relationship between languages. Towards this purpose, Gudschinsky (1956) outlines that a core tenet of Lexicostatistics lies on the idea of a group of lexical items being core or fundamental to a language and that these items change little over time. Furthermore Swadesh lists attempt to exclude loan words, being not considered part of the basic vocabulary. Hence with the content of the Swadesh list being designed to exclude varying items, it would be an inappropriate tool to examine contact. Although Lee and Sagart (2008) note that basic vocabulary is not resistant to borrowing. This is also seen in the import of Norse origin <they> into English's set of pronouns.

It would be perhaps helpful to disambiguate the nuances in Lexicostatistical methods. Swadesh lists being biased towards unvarying items does not rule out word-sense lists in general. Larger word-sense lists such as the IDS dictionaries might be a good resource for detecting loanwords due to French, given the large amount of entries and large amount of vocabulary topics. The Automated Similarity Judgment Program (ASJP; Holman et al. 2011) employs 50–100 word lists of items. These smaller Swadesh-type lists maybe inappropriate for assessing contact while larger lists might be more useful.

Lastly the lexicostatistical method employs word-to-word comparisons, whereby a word-sense's entries are cross-compared to the word-sense entries in the other language. Contact-induced change in the case of English was not merely in the lexicon but changes in orthography and morphology. French elements word-initial <z> and <v> and digraphs such as <ch> and <ou> were imported via French loanwords yet were abstracted and applied to other English words.

Compare OE spelling <cirice> to Modern English <church> or OE personal pronoun <ic> that began to be spelled <ich> in ME. A word-to-word analysis would fail to detect any French influence, such as comparing <church> to French <église>, only unless the orthography coincided, such as a comparison between English <chief> and Old French <chief>. Word-sense lists often ignore inflectional morphology, thus the changes in noun declensions or verb inflexions cannot be assessed internally in a language across time or between languages. Only derivational morphology might be detected with a word-to-word analysis. Yet only if the derivational affix coincided between the two languages as in English <ity>, Early Old French <itet>, Latin <itas>. Yet the same issue is present with neologisms. French origin affixes abstracted away from their imported context and were used to form new words. Interestingly word-sense lists (in practice) in ignoring inflectional morphology might obscure the similarity between words. Latin word-sense lists would primarily take the Nominative form, thus comparing Early Old French <itet> compared to Latin Nominative <itas> would obscure the closer relation to declined forms using <itat ...>.

Lexicostatistical methods and derivations thereof, such as Automatic cognate detection, if they use word-to-word comparisons would be limited by the words in their comparison set. The cognacy judgement can only be made between what is fed into the system. If a foreign origin word is not featured in the foreign dictionary (out-of-vocabulary) it cannot be detect. A comprehensive dictionary would be needed, such as the Anglo Norman Dictionary (Rothwell and Trotter 2008). Yet for all contexts a thorough dictionary will not be available. Yet a measure that could examine word parts could overcome this issue. French origin <connection> could be identified as French through the derivational suffix <tion> even if the dictionary did not contain the lexeme.

1.5 Character N-grams

A method was sought that would allow a sub-word analysis comparing word parts and one that could cross compare inventories of these word parts so as to this above stated French influence that would be missed by a word-to-word analysis. Character N-grams were chosen. According to Cavnar and Trenkle (1994), a character N-gram, as opposed to a word/term N-gram, is an N-sized character or letter slice of a word. For example, given the word <corpus>, a 2-character slice would produce the following collocations of letters, <_c>, <co>, <or>, <rp>, <pu>, <us>, <s_>, (where <_> represents a space).

Applied to a text sample, a frequency distribution of all appearing character co-occurrences can be computed. As Cavnar and Trenkle delineate, unigrams (N = 1) measure the frequency of the constituent alphabet of the language sample. Moving to bigrams (N = 2) to trigrams (N = 3), the character slices begin to capture the basic constituents of the language, comprising a language's pronouns, prepositions, affixes, determiners, etc. These character N-gram frequency profiles have been applied to text classification tasks. For instance, languages will have unique orthography or phonemes, as in accent marks in French, <*ñ*> in Spanish, or <þ> in Old English. Larger character slices begin to become more and more language specific (Tomović and Janičić 2007). These frequency profiles have been harnessed to identify the languages, such as the program TextCat (Feinerer et al. 2013), which successfully uses these N-gram profiles to classify the language of a written text sample with high accuracy. Rama et al. (2015) note that through comparing two languages' character N-gram frequency profiles, one can produce a single point distance numeric for the pair.

Within Lexicostatistics, the Levenshtein distance was used to compute pair-wise difference between features (in this case cognates) and thus derive a measure of similarity. Within Natural Language Processing 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 pair-wise; 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 normalized dot product of two normalized 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 totally similarity. The cosine similarity between two vectors *a* and *b* is as follows (Sidorov et al. 2014);

$$\cos ine(a,b) = \frac{\sum_{i=0}^{N} a_i b_i}{\sqrt{\sum_{i=0}^{N} a_i^2}\sqrt{\sum_{i=0}^{N} b_i^2}} \tag{1}$$

This method requires less knowledge of the language contained within the inputted text. Through this vector based method, all possible forms of features (letter collocations) are present in the vector and the forms in the language of comparison can co-occur and so be aligned allowing for the relevant distance between the languages to be calculated. Thus this method is alternative to top-down alignment of word forms or character sequences.

Character N-gram frequency profiles present the emergent pattern of a language, encapsulating the sum usage frequency of its component parts, its affixes, prepositions, pronouns, etc. Thus there is not only a focus on lexis but can measure morphology and orthographic usage.

However a major pitfall of this method is that they are quite biased to the text composition so that a large amount of one text type, per se a religious or a legal text, can skew measures of similarity between the texts. This is problematic in samples of low word count. In a diachronic study such as this, word count varies significantly between periods. So an overrepresentation of one text type in one period may inflate the degree of actual change, where the change is in fact due to the differing vocabulary within each text.

This methodology can be applied to multiple text samples and therefore allows for the study of diachronic variation through the use of text samples from different temporal epochs within the history of a language. Through use of a distance metric, similarity across time can be computed. Beyond language internal similarity, growths and declines in similarity over time between languages can be computed. This can allow for the possible study of the diversification of language families over time and contact between languages. Therefore it is posited that character N-gram models may provide an alternative to the study of language distance, synchronically and diachronically.

1.5.1 Background to the current study

Computational studies of historical languages that take a diachronic perspective have been few. Furthermore those who have assessed historical English often take whole linguistic epochs and seek no further subdivision. For instance, a dialectometry study, McMahon and Maguire (2011) employing word-sense lists to examine samples of historical English dialects (compared to samples of modern dialect), took samples from the gross epochs OE and ME yet did not make any distinction between Early or Late ME. Corpus studies, to the author's knowledge have been few. Some that have examined historical languages have a practical goal. Wahlberg et al. (2016) and Zampieri et al. (2016) used samples of diachronic corpora for temporal text classification, i.e. dating documents. In the case of Zampieri et al., word N-grams and morphosyntactic features were both employed in a historical corpus of Portuguese to classify test texts to a historical epoch. Similarly Wahlberg et al. employed a large corpus of Old Swedish and Latin documents. An image analysis of scribal handwriting was also combined with a character N-gram language model (from transcribed texts) to classify the date of the documents. The combined image and language model had the highest success in dating documents.

In contrast this current study aims at the measurement and quantification of change over time, within and between languages. Thus it aims to provide a quantitative confirmation of the contact between English and French. The current study is a largely exploratory study, aiming to examine the efficacy of character N-gram models to extract in a data-driven manner these orthographic, morphological, and lexical changes within the selected period of historical English outlined above. The hypothesis is accordingly broad but this is due to the lack of a previous quantitative pattern to compare the derived results with. Qualitative research has produced an expected trend of change within English and between English and French, with dates of expected changes to differing precision. Yet qualitative research can provide no indication of the magnitude of such changes. Ultimately similar enterprises have no quantitative trend to fit only an abstract idea. Lexicostatistical reconstructions of families' trees can only be graded on whether they fit what was expected or for cognate detectors that are only scorable by a qualitative evaluation. Thus this study refrains from more specific predictions, only examining the derived patterns in relation to these expected trends.

1.5.2 Debate over orthographic or phonetic space

Lexicostatistical studies have often transcribed their data into a phonetic transcription. However this presents a top-down augmentation of the data and furthermore an account of each languages orthographic to phonology correspondences must be known. In studies of several languages, an account of each study must be known. Secondly in diachronic examinations an account of the phonology of the language within each epoch must be known leading to a laborious effort. Firstly in large corpus studies this becomes a dubious practice where every spelling variation cannot be accurately transcribed and many top-down decisions must be made on what a symbol sequence represents. Secondly the accounts of phonological variation will not fit into predefined temporal epochs and often have fuzzy temporal boundaries as to when the change occurs. An application of a phonetic transcription will therefore have higher error at smaller temporal resolutions and produce artificial results whereby a pattern of increase or decrease in a variable will be in some part due to the augmentation of the data.

The decision of using a phonetic or orthographic transcription may depend on the goal. Lexicostatistics aims to detect genetic inheritance thus orthography may obscure this lineage. While in the case of change within English, orthographic change is one of the largest areas for change, also with the effect of French orthographic practices on English as in the case of <ou>. Thus similarly a phonetic transcription may obscure the detection of change towards another language.

This study primarily takes orthography as input data and assumes a common orthographic space such as that proposed by Singh and Surana (2007), assuming that the orthography of two languages in the same alphabet can be compared. It is important to acknowledge that alterations to the basic data have been made in transcription, with some corpora expanding scribal abbreviations or adding accent markers for French. However a common phonetic transcription is also examined in some cases. The ASJP has developed a phonetic transcription outlined in Brown et al. (2008). The ASJP code is perhaps apt to the application of historical languages. In several areas the phonetic information is reduced. For example in the ASJP code's transcription of vowels, vowels are transcribed roughly to a quadrant, thus high front vowels /i/ and /y/ are all represented by one symbol. Similarly in certain consonants voicing is ignored. This provides a strength for the transcription of historical languages, smoothing out contextual difficulties. Importantly for small changes in vowels between periods, the reduction to a larger grouping may then reduce the error.

1.5.3 Aims of the current study

The analyses fall into two main categories, intra-language analyses, i.e. examining the internal changes within English, and inter-language analyses, i.e. charting the rise and fall in similarity of English to several other historical languages across time. The languages in this case are the nearest Germanic relative, Frisian, and the Romance languages from which English borrowed, French (Anglo-Norman and Continental French) and Latin. Through this it is aimed to measure what Ellison and Miceli (2012) term contact-induced assimilation, changes in a language produced by contact with another that result in the one or both languages becoming more similar, in this case Anglo-Norman and English.

The use of character N-gram frequency profiles will be compared with a traditional lexicostatistical analysis. Swadesh lists for the gross epochs of historical English will be used and the ability to detect internal change and contact with Romance languages assessed. Furthermore the efficacy of

Swadesh lists within a linguistic epoch will be assessed compared to analyses using character N-gram profiles.

Based on the qualitative scholarship into English presented above, several expected outcomes have been delineated. They are as follows:

Language Internal Hypotheses:

1. For both the lexicostatistical and the character N-gram analyses, a change between OE and ME should be seen.
2. The N-gram analysis should reveal a large change from Early OE to Late ME. This should manifest as a growing decrease in similarity.
3. Within a chosen sub-period (in this case Early ME), both methods should detect the same pattern of internal change.
4. The most frequent words in the text corpus should change less between periods than less frequent words.
5. Top down defined closed-class words should change less across time as compared to open-class words.

Inter-Language Hypotheses:

1. An increase in similarity between Anglo-Norman and English should be seen for the ME period. Latin should have an increase but to a lesser degree.
2. For the N-gram analysis at a finer temporal resolution, this similarity with French should increase over time as per the increasing influx of loanwords.
3. French similarity to English closed-class words should be lower than for open-class words, as per Durkin's finding of a lack of change in common grammatical words.
4. Similarity between English and Frisian should decrease over time, as they grow apart due to English's contact with French.

2 Methods

2.1 General methodology

The main corpora and datasets used in this analysis will be outlined below. All other additional corpora and the subdivisions performed on them will be introduced where appropriate. For the temporal sub-periods delineated for each of these corpora and their respective word counts see Supplementary Materials.

2.1.1 Swadesh lists

The Swadesh-200 as defined by Millar and Trask (2015) was used. The data was drawn from the IDS (Key and Comrie 2015; Borin et al. 2013). The IDS provides rich word-sense lists for hundreds of languages often containing thousands of items. The lexical items for each language corresponding to the Swadesh-200 were used (the overlap was not complete, 196 items overlapped). Most lists present several variants with no information on which item is more frequent, thus one dominant headword was selected arbitrarily. The following languages were used, OE, ME, Latin, and Modern French in lieu of Old French. A Swadesh-200 list for Old Frisian was created from two sources. Primarily the incomplete list of Novotná and Blažek (2007) was used and supplemented with data from Bremmer (2009). In total, this list comprised 135 items. Alternatively, a transliteration into a phonetic alphabet (ASJP code) was created for all Swadesh lists (see Supplementary Materials)

2.1.2 Language corpora

Table 2 displays the text corpora used in this study. The primary source of historical English was the Helsinki Corpus of English texts. A breakdown of the temporal epochs of this corpus can be seen in Table 3. Each epoch is roughly 100 years in duration. Sentences of foreign languages annotated (such as Latin) in the Helsinki files were removed.

For an analysis of Early ME, variants of the Linguistic Atlas of Early Middle English (LAEME) were used. The parsed LAEME (P-LAEME) was used to create Swadesh lists for finer temporal periods within Early ME. P-LAEME consisted of a sub-section of the LAEME archive with word-sense tagging. This allowed for the creation of Swadesh lists for subperiods of ME. The P-LAEME files' attestation across time was assessed and three time bins were decided that had amble text content. They were as follows, 1225–1275, 1275–1300, 1300–1350. There was an attempt to create Swadesh-200 list for each period but surprisingly no period presented the necessary words to create a full list. This demonstrates that words that are considered to be core and unchanging need not necessarily appear often, even considering the large amount of data present in the P-LAEME files. Instead 405 lexical items found to present in each of the three periods were used. As each item presented multiple spellings, the most frequent spelling was chosen.

The Romance and Italic languages known to have contact with English were employed, them being Anglo-Norman French, Continental French, and

Table 2: Language corpora.

Language	Corpus	Source	Time Span	
English	Helsinki corpus of english texts	Kytö (1996)	850–1710	1,472,008
Middle English	LAEME	Laing and Lass (2007)	1150–1350	572,798
Middle English	P-LAEME	Alcorn et al. (2018)	1150–1350	
Anglo-Norman French	Base de Français Médiéval (BFM) & Anglo-Norman Correspondence Corpus	Guillot et al. (2007); Ingham (2011)	1100–1415	495,369
Central French	Base de Français Médiéval (BFM)	Guillot et al. (2007)	800–1500	3,119,102
Medieval Latin	Patrologia Latina database	Chadwyck-Healey (1995).	700–1200	11,878,537
Historical Portuguese	Post Scriptum database	Vaamonde et al. (2014), CLUL (2014)	1500–1600	81,792
Historical Spanish	Post Scriptum database	Vaamonde et al. (2014), CLUL (2014)	1500–1600	116,521
Old Frisian	Fryske akadamey's integrated scientific Frisan language database, codex Unia & TITUS corpus of old East Frisian Texts	Versloot and Nijdam (2011); Sytsema et al. (2012); Shannon (1999).	1300–1500	134,472
Basque	Basque-Spanish EiTB corpus of aligned comparable sentences	Etchegoyhen et al. (2016)	Modern	7,934,240
Hungarian	"*A Kis herceg*" - edition of "*Le Petit Prince*"	György (1993)	Modern	11,838
Lithuanian	"*Mažasis Princas*" - Lithuanian edition of "*Le Petit Prince*"	Kauneckas (1998)	Modern	10,944
Maori	The Legal Maori Archive	Darwin and Stephens (2009)	<1910	4,469,044

Medieval Latin. French text was drawn from the Base de Français Médiéval (BFM) with the Anglo-Norman Correspondence Corpus providing material between late 13th and the early fourteenth centuries where the BFM lacked material. For Medieval Latin, the Patrologia Latina database was used. Four volumes from each century in the database, from the 8th to the 13th, were randomly selected to provide an adequate representation of the language. In order to compare the sensitivity of the method to tease apart contact from

Table 3: Helsinki Corpus of English periodisation with word counts.

Helsinki epoch	Time period	No. of texts	
O1	<850	10	2,107
O2	850–950	21	88,507
O3	950–1050	92	245,115
O4	1050–1150	28	65,192
M1	1150–1250	30	107,277
M2	1250–1350	20	91,691
M3	1350–1420	45	171,590
M4	1420–1500	51	201,499
E.Mod.1	1500–1570	56	176,352
E.Mod.2	1570–1640	50	171,120
E.Mod.3	1640–1710	52	151,558

*O = Old, M = Middle, E.Mod = Early Modern.

different members of the same language family, historical Portuguese and Spanish were used. The aim was to assess whether the method could detect Anglo-Norman as being most similar to English than these close linguistic relatives that are also Romance languages.

In order to examine the accuracy in detecting differentiation of languages within the same language family from one another across time, samples of Old Frisian were compiled. Old Frisian is determined to be, in the literature, Old English's closest linguistic neighbour (Robinson 2005; Brinton and Arnovick 2006). The Frisian samples were often mixed with Latin. Only the subsection of the corpus, an edition of the Codex Unia, had Latin annotated and was thus stripped (see Section 3.4).

Measurements of the change in similarity would be better interpreted with a baseline to compare with. Control conditions were desired to provide an indication as to what magnitude English can be quantified as similar to a known-unrelated language. Basque was chosen as a control condition. Trask (1995) affirms that to date no credible evidence has been provided of a genetic relationship between Basque and any other Indo-European language. However Basque has been in heavy contact with its Romance neighbours; Spanish, Catalan, etc. This could lead to Romance elements being frequently present in the text corpus. Another comparison was added, a European language not from the Indo-European family, Hungarian, a Finno-Ugric language. Thirdly Lithuanian, an Indo-European language that is more distantly related to English than the other Romance languages, was used as it is related but did not have contact with English nor Romance languages. Both text samples were taken from translations of "*Le Petit Prince*" (de Saint-Exupéry 1948). Although Hungarian and Lithuanian are less apt languages orthographically speaking as they employ many letters not shared by English. Basque on the other hand mainly uses the Roman alphabet. Finally a non-European language that mainly employs the Roman alphabet was used, that could ensure no Romance contact. Maori was chosen as this control comparison. The Legal Corpus of Maori provided removal of any English words appearing in their texts, providing a clean sample.

2.1.3 String distance measure

A normalized Levenshtein distance (LDN) was employed to compute pairwise string distance/similarity between lexical items in the Swadesh-200 lists. The LDN is the Levenshtein distance between two words normalized by the maximum length of the two words (Rama et al. 2015), producing a value between 0 and 1.

2.1.4 N-gram processing

Character N-grams counts were generated using the R package Quanteda (Benoit and Nulty 2013). Prefix and suffix character N-grams were generated using the package Tau (Buchta et al. 2017). Each HC text sample's N-gram frequency profile was generated for unigrams, bigrams, and trigrams.

For the samples of historical English, an average frequency profile for each period was used. As attestation over time and geography varies greatly within each period, with some periods more greatly represented by West Saxon in OE, others by West Midlands dialect in ME, an average profile of all texts in an epoch was created to provide a representative profile of the whole period. The average frequency profile for each epoch was generated through generating a profile for each individual file from a period and summing the counts of each feature across files and dividing by the total number of files in the epoch. It was hoped that this would provide a better signature of the epoch, reducing the effect of a word count bias.

Each N-gram type was analysed both separately and in a combined model (N = 1–3; labeled below for descriptive purposes as "All N-grams"). The lowest occurring features were filtered and summarized in an 'informational tail' which contained data on the sum count of the items removed, the mean and median value, and the standard deviation of the filtered items, thereby allowing removed features to contribute to the profile.

2.2 Chi-square statistic

In order to examine the differences between N-grams across periods, the Chi-Square (X^2) Test of Independence was employed. The Chi Square test is a possible method of feature selection employed in Natural Language Processing. Fuka and Hanka (2001) note that the Chi Square test gave good results and should not be ignored. Kilgarriff (2001) discussed the application of Chi Square contingency tables to measure how similar one corpus/text was to another, measuring differences both positive and negative between them (in his case word features). Kilgarriff concluded that Chi Square is a suitable measure for comparing corpora.

A feature selection technique was favored over a feature reduction technique like Principle Component Analysis (PCA) so that features of change could be isolated. PCA and neural network approaches have a property of opacity whereby the components that discriminate categories are not immediately interpretable. Studies like Zampieri et al. (2016) used classification tasks to highlight features of change between temporal periods (in Portuguese). Classification methods such as Random Forest can be used for feature selection, highlighting variable importance. Variable importance can highlight which features contribute to the accurate discrimination of two (or more) classes (Liaw and Wiener 2002; Kuhn 2012). Yet a variable that is predicative of a class does not necessary capture all or some variables that change between classes. A variable could decrease in frequency between two English epochs yet not be predicative for discriminating two classes. Methods, such as the Chi Square test, were sought that enabled inspection of the information associated with individual features, measuring whether they increase or decrease in frequency. We view the test as appropriately simple in its assumptions, and therefore more widely viable for the purpose of comparisons that were sought to be made.

The Chi Square test calculates the probability of the difference between two categorical variables being due to random chance. Pairwise comparisons between features in two chosen temporal epochs were carried out. Each Chi-Square test examined the raw frequency counts of a pair of features (an N-gram) across two periods, over the sum count of all other N-grams in the two select periods. Through this significant changes in frequency count could be detected between periods.

Chi Square Residuals measures the degree of difference between the variables, through the subtraction of the expected count of the feature, given no significant difference, from the actually observed count. The calculation of the Chi Square test is as follows;

$$\chi^2 = \sum \frac{(observed - \exp ected)^2}{\exp ected} \qquad (2)$$

Through this a measure of the difference between two feature counts can be achieved. This also allowed inspection of the magnitude of change of an N-grams frequency between periods.

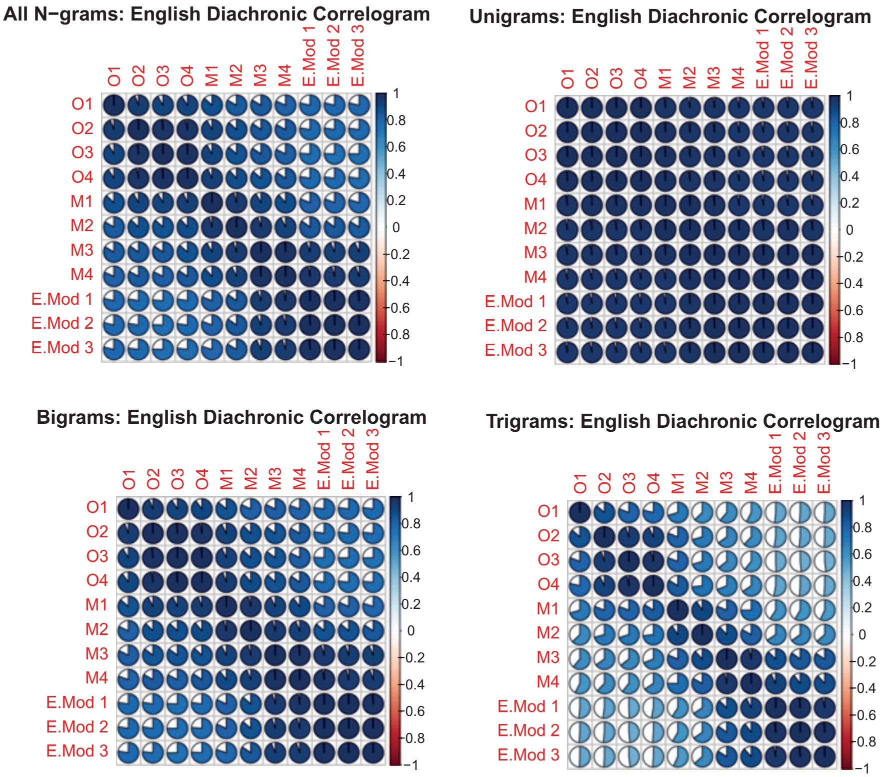
3 Results

3.1 English language internal analyses

Pairwise distances between each HC language epoch were computed. In Figure 1 the resulting correlation matrix is presented via a correlogram[1]. Unigrams, measuring the alphabet of the samples, shows only minor changes throughout the entire period of historical English from O1 to E.Mod.3, indicating that the usage each letter has change only slightly. However a change of ~1–2% is present possibly due to loss of early English letters such as thorn <þ> and eth <ð>. Bigrams, collocations of 2 letters, and trigrams, clusters of 3 letters, present growing decrease in shared features over time throughout the ME period.

Among the OE periods, there appears relative stability. Hogg (1992) views OE's orthographic usage as mostly stable across time and these results replicate this. From the earliest English, O1, there is a decline across the remainder of the OE period but this may be due to O1's small word count, due to data scarcity, preventing the full range of features to be present for comparison. As can be seen for unigrams, the alphabet usage frequency from O1 to O4 remains unchanged, indicating no major changes. Similarly the Early Modern period remains uniform, possibly detecting the standardization of English in this period.

[1] Correlograms are a visual depiction of a similarity matrix. In this case the degree of similarity between two periods is displayed in pie-graphs, with the amount of shaded area corresponding to the percentage of similarity. The degree of shading corresponds to this level of similarity. Each pie-graph demonstrates the similarity between the period on the y-axis to the aligned period on the x-axis. Periods have been aligned in the correct temporal sequence, with the arrow of time proceeding rightwards. For some similarity matrices where the differences in similarity are too small to be seen visually, numeric values have been used. In the below cases the entire matrix has been visualized to allow examination of each time point to every other time point but for some data only sections of the similarity matrices have been presented for ease of interpretation.


All N−grams: English Diachronic Correlogram

Unigrams: English Diachronic Correlogram

Bigrams: English Diachronic Correlogram

Trigrams: English Diachronic Correlogram

Figure 1: Diachronic English internal similarity correlograms.

3.2 Lexicostatistical measure of change

The change in similarity between the gross epochs OE and ME were measured through the sum LDN between word-senses in both languages using the Swadesh-200 lists. It was measured for both a phonetic and an orthographic transcription that there is a decrease in similarity between the two stages of English. In Figure 2 this substantial decrease in similarity can be seen, with around 40–50% decrease in similarity for both transcriptions. This indicates that the lexicon of OE has undergone a great change.

This is detectable in both an orthographic and a phonetic transcription. However it is shown that for orthography there is a larger decrease in similarity by 10%. This shows that English's orthography has changed more between the epochs than it had phonetically. This raises once more the question of the goal of the analysis. If it should desire to detect the differentiation of English, then an

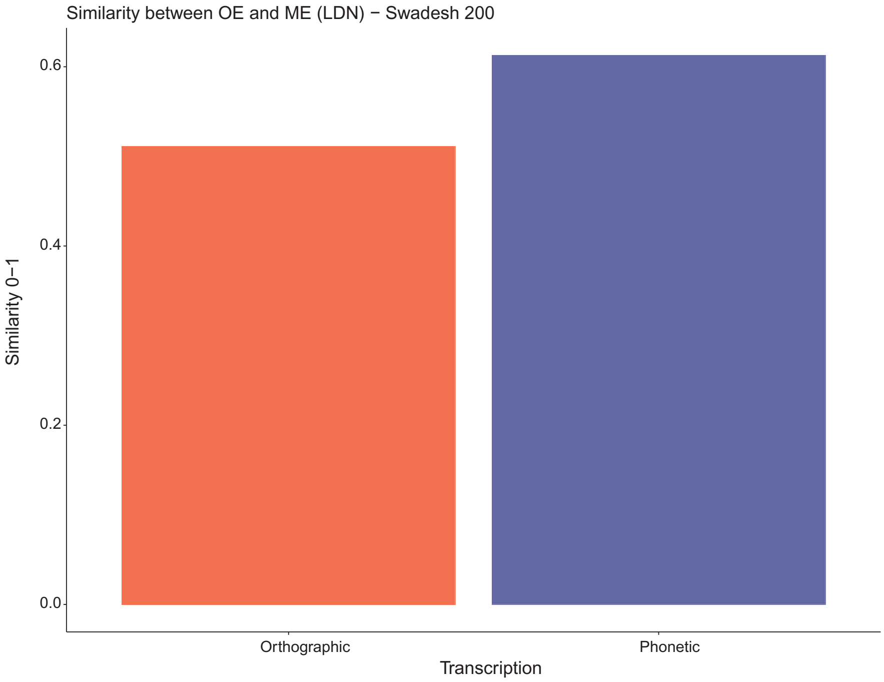


Figure 2: Relative similarity between OE and ME (LDN).

orthographic transcription measures this well while a phonetic transcription better captures the relation of ME to OE by showing a greater similarity.

The ability to assess linguistic contact through Swadesh list was assessed. Figure 3 demonstrates that the method was indeed sensitive enough to detect contact-induced change from French and Latin. It can be seen both French and Latin register an increase in similarity between OE and ME, with French demonstrating the largest increase. Interestingly French has a larger effect on English orthographically while the phonetic transcription displays a much smaller increase. This may be a by-product of the use a Modern French Swadesh list where a phonetic transcription renders many differences from the medieval variety English was contacted by. For instance modern French's unpronounced final consonants and its different pronunciation of <ch> and <j>, which at the time of contact would have been pronounced the same (Einhorn 1974). A phonetic transcription of Latin renders little increase perhaps indicating that the orthography shared by French might be causing the demonstrated increase in orthographic Latin similarity.

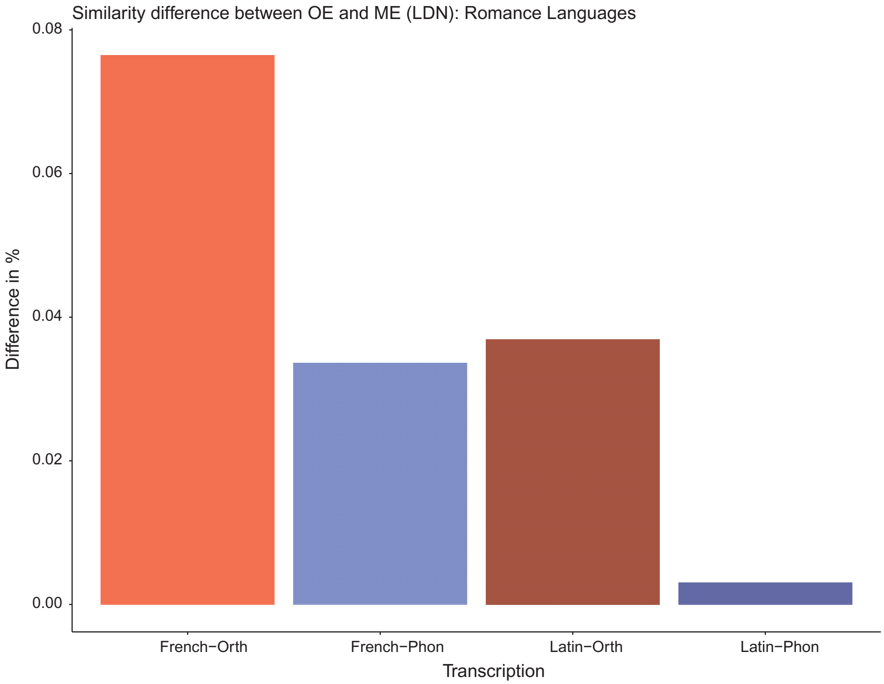


Figure 3: Increase in similarity of Romance languages to English (measured between OE and ME).

3.3 Early ME internal change

The internal change assessed by a word sense list within the delineated temporally fine sub-periods of Early Middle English is displayed in Figure 4. It can be seen that a moderate change occurred between 1275–1300 but that a slightly larger change occurs from 1300–1350. This demonstrates that a lexicostatistical method can detect change between periods yet however this is not a traditional Swadesh list but a larger vocabulary (see Section 2.1.1).

The ability of N-gram frequency profiles to derive the same patterns as top-down compiled lists was assessed in the corresponding corpus files of the LAEME archive (although there was not a complete correspondence between the two corpora). Figures 5 (a) and 5(b) demonstrate that N-gram profiles abstract the same relative pattern between the periods, with a change occurring between 1275 and 1300 and a larger change occurring between 1300–1350. This demonstrates that methods unguided by an expert can derive the same patterns of change.

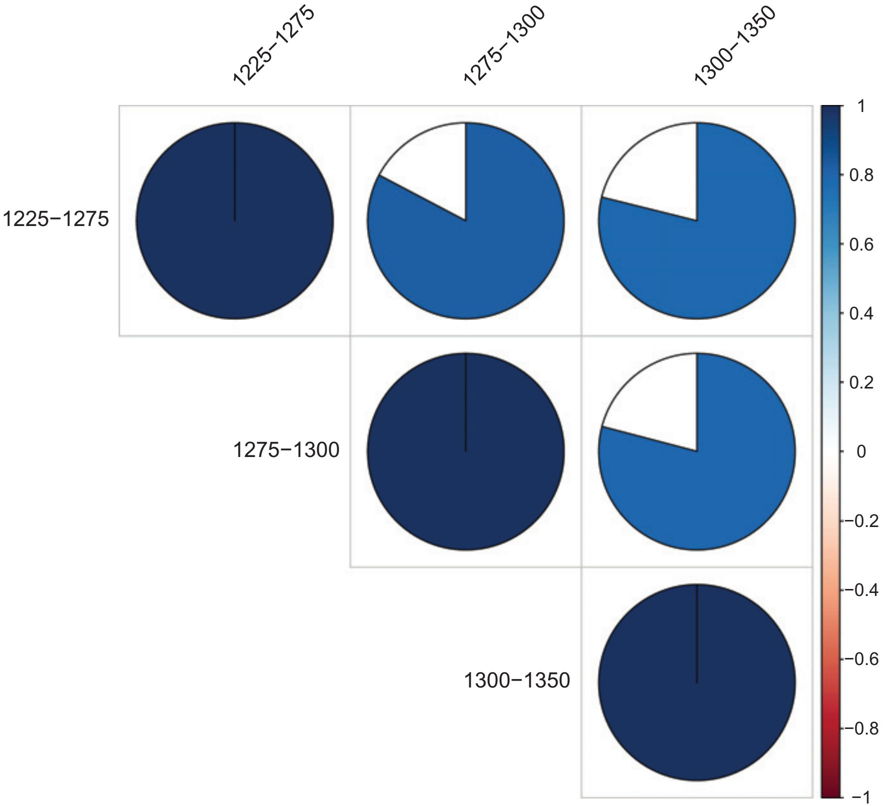


Figure 4: P-LAEME correlogram – measured by LDN.

3.4 Features of change in text corpora

The Chi Square test of independence was used to examine epoch-to-epoch changes in frequency for each feature, to examine where and in what direction, increase or decrease, the features varied. Each period was compared to the next temporally adjacent period from O4 through to E.Mod.1. Three of the important contrastive periods are displayed below. In Tables 4 and 5, the change from OE to Early ME is examined, comparing period O4 to M1. Secondly the change from Early ME to Late ME is examined through comparison of period M2 to M3, seen in Tables 6 and 7. Lastly the change from Late ME to Early Modern English is examined by comparison of M4 to E.Mod.1, seen in Tables 8 and 9. For each contrast, the top increasing and decreasing features are presented, 5 unigrams, 10 bigrams, and 15 trigrams.

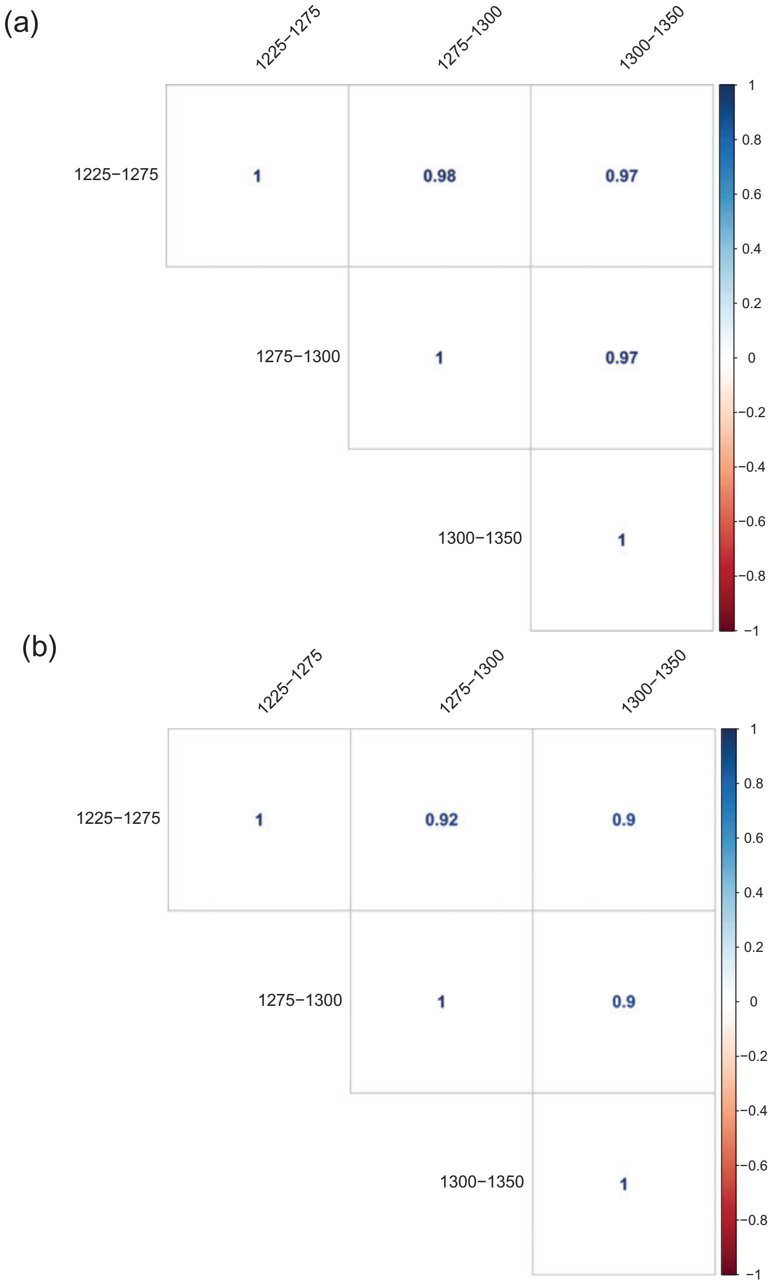


Figure 5: (a) LAEME 2-gram correlogram – measured by cosine similarity. (b) LAEME 3-gram correlogram – measured by cosine similarity.

Table 4: O4 > M1 – Changing features (derived via X^2).

N-gram	*Residual*	*P value*	X^2	*N-gram*	*Residual*	*P value*	X^2
1gram				1gram			
Increasing				Decreasing			
ʒ	**30.6**	*0.00*	*2452*	g	**–40.7**	*0.00*	*4390*
e	**20.8**	*0.00*	*1292*	æ	**–36.5**	*0.00*	*3511*
ꝥ*	**20.6**	*0.00*	*1111*	y	**–31.4**	*0.00*	*2580*
k	**19.4**	*0.00*	*984*	a	**–19.5**	*0.00*	*1059*
h	**18.25**	*0.00*	*900*	c	**–12.3**	*0.00*	*403*
2gram				2gram			
Increasing				Decreasing			
ch	**26.2**	*0.00*	*1792*	ge	**–37.9**	*0.00*	*3769*
ue	**24.6**	*0.00*	*1572*	þæ	**–32.2**	*0.00*	*2711*
en	**24.4**	*0.00*	*1574*	_g	**–29.6**	*0.00*	*2301*
ʒe	**23.5**	*0.00*	*1445*	æt	**–29.2**	*0.00*	*2229*
þ_*	**20.3**	*0.00*	*1072*	an	**–22.7**	*0.00*	*1370*
_þ*	**19.9**	*0.00*	*1031*	a_	**–20.6**	*0.00*	*1114*
_ʒ	**19.8**	*0.00*	*1017*	um	**–20**	*0.00*	*1048*
ei	**19.2**	*0.00*	*959*	cy	**–18.3**	*0.00*	*871*
ke	**17.4**	*0.00*	*783*	ær	**–18**	*0.00*	*846*
_i	**16.7**	*0.00*	*727*	ra	**–18**	*0.00*	*845*
3gram				3gram			
Increasing				Increasing			
en_	**27.5**	*0.00*	*1984*	_ge	**–40.2**	*0.00*	*4241*
che	**20.5**	*0.00*	*1098*	_þæ	**–32.1**	*0.00*	*2691*
_þ_*	**19.5**	*0.00*	*992*	an_	**–29.4**	*0.00*	*2324*
_al	**18.9**	*0.00*	*937*	þæt	**–29.3**	*0.00*	*2250*
_ht	**18**	*0.00*	*847*	um_	**–26**	*0.00*	*1769*
ich	**16.9**	*0.00*	*751*	on_	**–21**	*0.00*	*1165*
_ʒe	**16.5**	*0.00*	*710*	eal	**–16.5**	*0.00*	*731*
ht_	**15.8**	*0.00*	*653*	þam	**–16.5**	*0.00*	*709*
þat	**15.7**	*0.00*	*641*	_cy	**–16.3**	*0.00*	*692*
eð_	**15.1**	*0.00*	*597*	_ea	**–16.3**	*0.00*	*693*
ch_	**14.8**	*0.00*	*574*	hi_	**–16.2**	*0.00*	*688*
ss_	**14.65**	*0.00*	*558*	cyn	**–16.1**	*0.00*	*676*
aue	**14.2**	*0.00*	*529*	ra_	**15.7**	*0.00*	*643*
_ha	**14**	*0.00*	*511*	æs_	**–15**	*0.00*	*589*
uer	**12.9**	*0.00*	*432*	_wæ	**–14.9**	*0.00*	*583*

*Crossed thorn

Table 5: O4 > M1 – Co-appearing features.

Increasing features:				Decreasing features:			
N-gram	*Co-appearing feature*	*Residual*	X^2	*N-gram*	*Co-appearing feature*	*Residual*	X^2
1-grams							
ʒ				g			
	ʒe	23	*1441*		ge	−37	*3791*
	_ʒ	20	*1072*		_g	−31	*2459*
	iʒ	11	*340*		æg	−17	*728*
e				æ			
	ue	24	*1564*		þæ	−32	*2722*
	en	24	*1547*		æt	−29	*2239*
	ʒe	23	*1441*		ær	−18	*852*
þ*				y			
	_þ	21	*1103*		cy	−18	*873*
	ø				yn	−18	*836*
	ø				yr	−15	*836*
2-grams							
ch				ge			
	che_	16	*647*		e_ge	−20	*1077*
	_ich	11	*334*		s_ge	−15	*593*
	lich	11	*313*		n_ge	−14	*516*
ue				þæ			
	eaue	9	*199*		þæt_	−29	*2255*
	uer_	9	*199*		e_þæ	−17	*781*
	eoue	8	*168*		n_þæ	−14	*503*
en				_g			
	en_a	14	*517*		e_ge	−20	*1077*
	en_h	12	*366*		_ges	−16	*691*
	enn_	11	*351*		s_ge	−15	*593*
3-grams							
en_				_ge			
	en_and	10	*257*		and_ge	−9	*219*
	þenne_	6	*98*		secge_	−9	*207*
	eiden_	6	*80*		_georn	−9	*197*
che				_þæ			
	_muche	7	*116*		_þæt_h	−18	*852*
	eliche	5	*69*		_þære_	−9	*201*
	_riche	5	*52*		and_þæ	−8	*189*
_þ_*				an_			
	_þ_he_	7	*112*		an_and	−13	*425*
	_þ_is_	6	*80*		yððan_	−7	*129*
	_þ_ha_	5	*66*		ian_an	−7	*128*

Table 6: M2 > M3 – Changing features (derived via X^2).

N-gram	***Residual***	***P value***	***X^2***	***N-gram***	***Residual***	***P value***	***X^2***
1gram				1gram			
Increasing				Decreasing			
t	**25.2**	*0.00*	*1978*	þ	**–32**	*0.00*	*3075*
y	**21.6**	*0.00*	*1397*	ð	**–23.5**	*0.00*	*1615*
h	**11.8**	*0.00*	*428*	ȝ	**–17.4**	*0.00*	*891*
p	**10.5**	*0.00*	*325*	z	**–15.8**	*0.00*	*730*
a	**10.4**	*0.00*	*335*	e	**–12.6**	*0.00*	*557*
2gram				2gram			
Increasing				Decreasing			
th	**47.3**	*0.00*	*6638*	þe	**–26.6**	*0.00*	*2098*
_t	**31.4**	*0.00*	*2933*	_h	**–24.1**	*0.00*	*1731*
gh	**16.7**	*0.00*	*815*	_þ	**–19.8**	*0.00*	*1315*
yn	**16.3**	*0.00*	*779*	eþ	**–17.7**	*0.00*	*919*
ee	**16.1**	*0.00*	*761*	_z	**–17.5**	*0.00*	*894*
ve	**15.4**	*0.00*	*191*	ȝt	**–17.16**	*0.00*	*860*
ly	**14.4**	*0.00*	*606*	iȝ	**–15.8**	*0.00*	*733*
wh	**13.9**	*0.00*	*564*	þo	**–15.6**	*0.00*	*718*
ha	**13.8**	*0.00*	*567*	e_	**–15**	*0.00*	*701*
oo	**13.5**	*0.00*	*532*	de	**–14.3**	*0.00*	*611*
3gram				3gram			
Increasing				Increasing			
_th	**44**	*0.00*	*5709*	þe_	**–23.2**	*0.00*	*1590*
the	**34.4**	*0.00*	*3487*	þet	**–22.9**	*0.00*	*1534*
tha	**22**	*0.00*	*1426*	_þe	**–21**	*0.00*	*1297*
hat	**20.8**	*0.00*	*1271*	_ic	**–18.5**	*0.00*	*1001*
thi	**16.8**	*0.00*	*828*	et_	**–18.3**	*0.00*	*981*
ght	**15.2**	*0.00*	*678*	eþ_	**–17**	*0.00*	*844*
f_t	**14.1**	*0.00*	*583*	_hi	**–16.9**	*0.00*	*838*
_wh	**14.1**	*0.00*	*525*	e_h	**–16.6**	*0.00*	*811*
ly_	**12.8**	*0.00*	*478*	e_þ	**–16.3**	*0.00*	*782*
eth	**12.2**	*0.00*	*438*	hij	**14.2**	*0.00*	*590*
he_	**12**	*0.00*	*425*	iȝt	**–14.1**	*0.00*	*579*
sch	**11.8**	*0.00*	*407*	_he	**–14**	*0.00*	*580*
not	**11.4**	*0.00*	*382*	_þo	**–14**	*0.00*	*573*
e_t	**11.3**	*0.00*	*379*	_ht	**–13.9**	*0.00*	*565*
tho	**11.3**	*0.00*	*377*	ne_	**–13.5**	*0.00*	*541*

Table 7: M2 > M3 – Co-appearing features.

Increasing features:				**Decreasing features:**			
N-gram	***Co-appearing feature***	***Residual***	***X^2***	***N-gram***	***Co-appearing feature***	***Residual***	***X^2***
1-grams							
t				**þ**			
	th	**47**	*6662*		þe	**−27**	*2100*
	_t	**33**	*3180*		_þ	**−22**	*1489*
	it	**13**	*476*		þ_	**−19**	*1104*
y				**ð**			
	yn	**16**	*778*		ð_	**−14**	*602*
	ly	**14**	*607*		_ð	**−14**	*602*
	ey	**13**	*462*		ðe	**−13**	*500*
h				**ȝ**			
	th	**47**	*6662*		ȝt	**−17**	*863*
	gh	**17**	*870*		iȝ	**−16**	*735*
	wh	**13**	*564*		oȝ	**−13**	*506*
2-grams							
th				**þe**			
	the_	**30**	*2576*		e_þe	**−15**	*638*
	e_th	**21**	*1267*		þe_h	**−10**	*289*
	that	**21**	*1238*		o_þe	**−9**	*231*
_t				**_h**			
	_the	**33**	*3182*		_hij	**−14**	*591*
	_tha	**22**	*1457*		_ht_	**−14**	*565*
	hat_	**21**	*1271*				
gh				**_þ**			
	ght_	**13**	*514*		_þe_	**−19**	*1014*
	nogh	**6**	*106*		e_þe	**−15**	*638*
	ghte	**6**	*95*		_þo_	**−13**	*491*
3-grams							
_th				**þe_**			
	that	**21**	*1237*		eþ_þe_	**−8**	*205*
	_of_th	**15**	*631*		_þe_er	**−7**	*141*
	of_the	**13**	*460*		_þe_he	**−7**	*141*
the				**þet**			
	of_the	**13**	*460*		þet_is	**−9**	*256*
	nd_the	**9**	*233*		þet_he	**−8**	*164*
	in_the	**9**	*212*		þet_hi	**−7**	*128*
tha				**_þe**			
	that_t	**9**	*244*		eþ_þe_	**−8**	*205*
	that_i	**8**	*184*		_þe_er	**−7**	*141*
	that_h	**7**	*145*		_þe_he	**−7**	*141*

Table 8: M4 > E.Mod 1 – changing features (derived via X^2).

N-gram	*Residual*	*P value*	*X^2*	*N-gram*	*Residual*	*P value*	*X^2*
1gram				1gram			
Increasing				Decreasing			
t	**22.9**	*0.00*	*1068*	þ	**−76.6**	*0.00*	*11,194*
h	**19.8**	*0.00*	*788*	y	**−32.1**	*0.00*	*2031*
i	**12.6**	*0.00*	*313*	ȝ	**−28.5**	*0.00*	*1536*
u	**10.2**	*0.00*	*204*	l	**−11**	*0.00*	*244*
~	**7.7**	*0.00*	*113*	z	**−7.4**	*0.00*	*103*
2gram				2gram			
Increasing				Decreasing			
th	**39.6**	*0.00*	*3045*	_þ	**−71.1**	*0.00*	*9643*
ea	**31.5**	*0.00*	*1888*	þe	**−55**	*0.00*	*5747*
_t	**30.4**	*0.00*	*1802*	þa	**−39.3**	*0.00*	*2926*
he	**27.4**	*0.00*	*1454*	yn	**−27.8**	*0.00*	*1480*
in	**20.5**	*0.00*	*802*	þi	**−22.7**	*0.00*	*978*
ie	**20**	*0.00*	*760*	_ȝ	**−22.6**	*0.00*	*970*
mi	**16.6**	*0.00*	*522*	ys	**−22.2**	*0.00*	*935*
dg	**15.6**	*0.00*	*463*	ht	**−20.1**	*0.00*	*768*
im	**14.9**	*0.00*	*421*	ȝe	**−19.1**	*0.00*	*699*
ni	**13.6**	*0.00*	*287*	sc	**−19.1**	*0.00*	*697*
3gram				3gram			
Increasing				Increasing			
the	**43.4**	*0.00*	*3625*	_þe	**−52.1**	*0.00*	*5152*
_th	**37.8**	*0.00*	*2765*	þe_	**−46.3**	*0.00*	*4064*
ing	**28.1**	*0.00*	*1500*	_þa	**−38.7**	*0.00*	*2836*
ine	**19.3**	*0.00*	*703*	þat	**−34.7**	*0.00*	*2279*
ear	**18.6**	*0.00*	*658*	e_þ	**−33.2**	*0.00*	*2092*
he_	**18**	*0.00*	*623*	sch	**−24.8**	*0.00*	*1172*
her	**17.7**	*0.00*	*601*	n_þ	**−24.5**	*0.00*	*1136*
ie_	**17.4**	*0.00*	*576*	d_þ	**−23.5**	*0.00*	*1049*
tio	**16.5**	*0.00*	*514*	yn_	**−23.6**	*0.00*	*1023*
ed_	**15.9**	*0.00*	*484*	_þi	**−21.8**	*0.00*	*901*
rea	**15.8**	*0.00*	*472*	_ht	**−21.7**	*0.00*	*897*
hea	**15.2**	*0.00*	*437*	þer	**−21.1**	*0.00*	*844*
hin	**15.2**	*0.00*	*438*	f_þ	**−21.0**	*0.00*	*836*
you	**15.1**	*0.00*	*434*	t_þ	**−20.7**	*0.00*	*813*
ain	**15.1**	*0.00*	*432*	s_þ	**−20.3**	*0.00*	*778*

Table 9: M4 > E.Mod 1 – Co-appearing features.

Increasing features:				**Decreasing features:**			
N-gram	***Co-appearing feature***	***Residual***	X^2	***N-gram***	***Co-appearing feature***	***Residual***	X^2
1-grams							
t				**þ**			
	th	**40**	*3043*		_þ	**−72**	*9907*
	_t	**31**	*1813*		þe	**−55**	*5748*
	ot	**13**	*318*		þa	**−39**	*2925*
h				**y**			
	th	**40**	*3043*		yn	**−28**	*1482*
	he	**27**	*1454*		ys	**−22**	*953*
	h_	**13**	*330*		ym	**−15**	*433*
i				**ȝ**			
	in	**21**	*802*		_ȝ	**−23**	*994*
	ie	**20**	*763*		ȝe	**−19**	*700*
	mi	**17**	*547*		ȝt	**−13**	*310*
2-grams							
th				**_þ**			
	the_	**30**	*1678*		_þe_	**−47**	*4140*
	ther	**23**	*1046*		_þat	**−35**	*2280*
	them	**16**	*476*		e_þe	**−22**	*927*
ea				**þe**			
	grea	**15**	*401*		e_þe	**−22**	*927*
	eare	**14**	*389*		n_þe	**−19**	*704*
	eate	**14**	*371*		þe_s	**−19**	*696*
_t				**þa**			
	_the	**41**	*3293*		þat_	**−35**	*2277*
	of_t	**15**	*443*		e_þa	**−21**	*859*
	f_th	**15**	*439*		s_þa	**−13**	*325*
3-grams							
the				**_þe**			
	of_the	**17**	*551*		_of_þe	**−18**	*594*
	to_the	**13**	*342*		of_þe_	**−17**	*567*
	nd_the	**12**	*272*		and_þe	**−15**	*409*
_th				**þe_**			
	of_the	**17**	*551*		of_þe_	**−17**	*567*
	them	**16**	*476*		to_þe_	**−14**	*368*
	_of_th	**15**	*450*		in_þe_	**−13**	*335*
ing				**_þa**			
	_thing	**13**	*299*		_þat_i	**−14**	*355*
	ing_th	**12**	*281*		_þat_h	**−13**	*325*
	ing_of	**10**	*197*		_þat_þ	**−13**	*321*

To further elucidate the meaning of these varying features, co-appearing features were generated. N-gram slices appearing directly adjacent to the feature of interest were derived. For instance, a highlighted bigram, <th> co-appearing next to the bigram <at> to elucidate the word possibly being <that>. Co-appearing features with the N-grams of interest were measured and their period-to-period frequency changes were measured by Chi Square. Due to space limitations, only the top 3 features of interest per N-gram size will be shown as an illustration.

3.5 O4 compared to M1

From O4 to M1, an increase of yogh and crossed thorn can be seen, as well as <e>. Yogh appears to be carried by <ʒe>, a spelling variant of <ge> the prefix demarking an OE past participle. It can be seen that <ge> concurrently decreases. Co-appearing N-grams (Table 5) reveal that this increase in <e> is related to <en_>, the ME verb infinitive marker. It can be seen that <an_> the OE verb infinitive marker decreases. Co-appearing features further show this might indeed be verb infinitive suffix decreases, showing <an_and> and <ian_an>, two word final suffices coinciding with the word <and>.

French origin orthography has been detected, such as <k> and <ch>. <ch> can be seen to now be used to represent /ʃ/, as in seen in the rise of <lich> for OE <lic>, <riche> for OE <rice> and the personal pronoun <_ich>, formerly spelled <ic>.

OE symbols such as ash see a decrease. Co-appearing features show that <þæt> saw a large decrease in this period, while the spelling variant <þat> sees an increase. The decrease of <_wæ> and <æs_>, most likely related to past tense of the verb “to be”, <wæs>, possibly indicates a decrease in the use of ash generally.

Some evidence of the decline of the dative case can be seen with the reduction in the frequency of dative plural and dative adjective suffix <um_> and the dative plural determiner <þam>.

3.6 M2 compared to M3

The comparison of Early to Late ME, reveals the period in which orthographic symbols such as <þ>, <ð>, and <ʒ> see a large decrease in usage, and their replacements see an increase in usage, namely <th> and <gh>. It can be seen that determiners such as <þe> and <þet> decrease while <the> and <that> increase.

Yogh decreases as <gh> rises. Spellings such as <iȝ> and <ȝt> and <oȝ> drop in frequency, while <gh> now appears to replace these spelling with <ght_> and <nogh>.

This method demonstrates the ability to pinpoint in time the period in which certain spellings occur. Vowel variants <oo> and <ee> show an increase in this period, allowing for the detection of their major use. Similarly the consonant cluster <sch> is detected to have a large rise in usage, as outlined by Freeborn (2006) to have occurred between the twelfth to fourteenth centuries.

3.7 M4 compared to E.Mod.1

The contrast of M4 to E.Mod.1 shows the ME symbols <þ> and <ȝ> fall further from usage, with thorn falling the sharpest. <sch>, which rose in Late ME, in this period falls from usage. Many ubiquitous elements of Modern English are detected to have increased in this contrast. <th> spellings further rise as thorn variants decrease across the board. Co-appearing features (Table 9) mainly display <th> appears with the word <the> and interesting <_them_> the 3rd person plural pronoun.

The character slice <ing>, increasing in this period, possibly indicates the verbal progressive marker. Co-appearing features reveal this <ing> is a word final slice, so indeed could be the rise of the progressive verb ending. Interestingly a rise in he spelling <_thing_> is registered. This might have risen along with <ing> due to the rise of <th> spellings also. The verb past tense marker <ed_> can be seen. Spellings of pronouns such as <you>, <he>, and <her> rise in this period. Interestingly the digraph <ea> rises in frequency. It is unclear what is the cause of this. Large sized character slices would be needed to elucidate this but co-appearing features might suggest it is spellings such as <great>, <eat>, or <ear>.

3.8 Inter-language comparisons

To test whether similarity between N-gram profiles could detect the predicted growth in similarity due to gross lexical influx, similarity between English per epoch and the Romance Languages and the Control condition, Basque were computed. The period of examination began with the baseline period of O4 through to the end of the Early Modern period.

Figure 6 displays the trend in similarity to English per language. Both varieties of French increase in similarity to English from the end of OE to M1

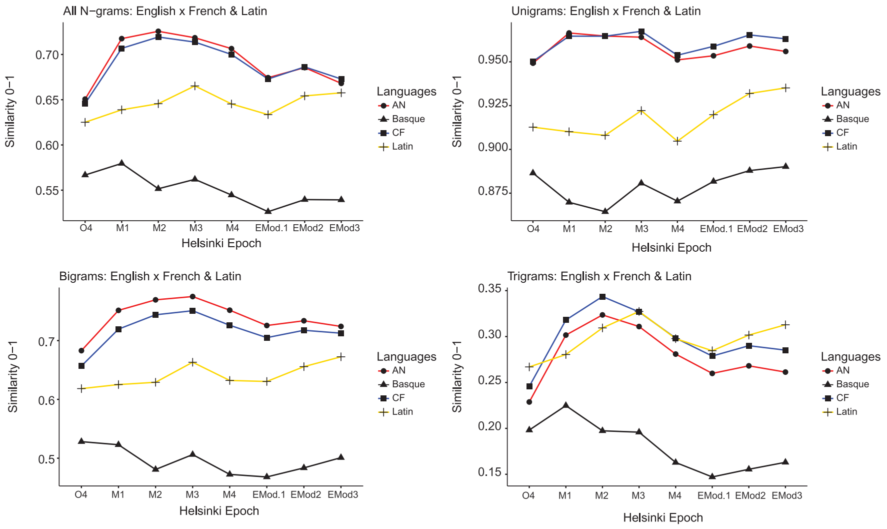


Figure 6: Diachronic similarity between English and Romance languages.

as expected. A similar trend for French is apparent in all N-gram slice sizes. Similarity continues to rise with trigrams displaying a sharper rise at M2. Similarity begins to fall between M4 and E.Mod.1. However a rise in similarity between French and English begins again in the later Early Modern period. It is not clear if a difference between the varieties of French can be discerned, with variance across N-gram sizes as to the degree of separation between them.

Latin, in contrast, does not show a steep rise as with French, demonstrating that Latin and French can be discriminated in their influence on English. Furthermore Latin at M3 shows a sharp increase in similarity to English. This corresponds well with the period of lexical influx outlined by Durkin (2014) as occurring between 1350–1450, with period M3 covering 1350–1420. Latin sees a second growth in the Early Modern period, which corresponds to the influx of Latin in the 1600s (Durkin 2014). This indicates that the measure was sensitive in detecting different periods of linguistic borrowing.

The control condition of Basque performed well setting a relative limit of similarity due to chance resemblance. All Romance languages were above this lower limit. However the similarity trend is not stable, signaling caution in the interpretation of increases and decreases in similarity. As English changes, it can thus increase or decrease the probability of chance resemblances with unrelated languages. This sample of Modern Basque will undoubtedly have influence from Spanish being languages in close proximity.

Lastly it is important to note that the cosine similarity measure is a relative measure and it cannot be concluded from the measure alone that it was not French or Latin that became more similar to English. Taking into account the greater internal change of English in comparison to the internal change of French can provide empirical evidence that English was the likely candidate of the change in similarity between the languages.

3.9 Anglo-Norman compared to other Romance languages

To tease apart whether this method can isolate the contact language as Anglo-Norman, a comparison of the increase of French to related Romance languages was computer. Figure 7 demonstrates that Anglo-Norman demonstrates a higher increase compared to historical Spanish and Portuguese. While Anglo-Norman also demonstrates a higher baseline similarity to OE, the increase is larger than the increases demonstrated by other Romance languages. Spanish and Portuguese also demonstrate an increase that may be due to the genetic relation of French to these sister

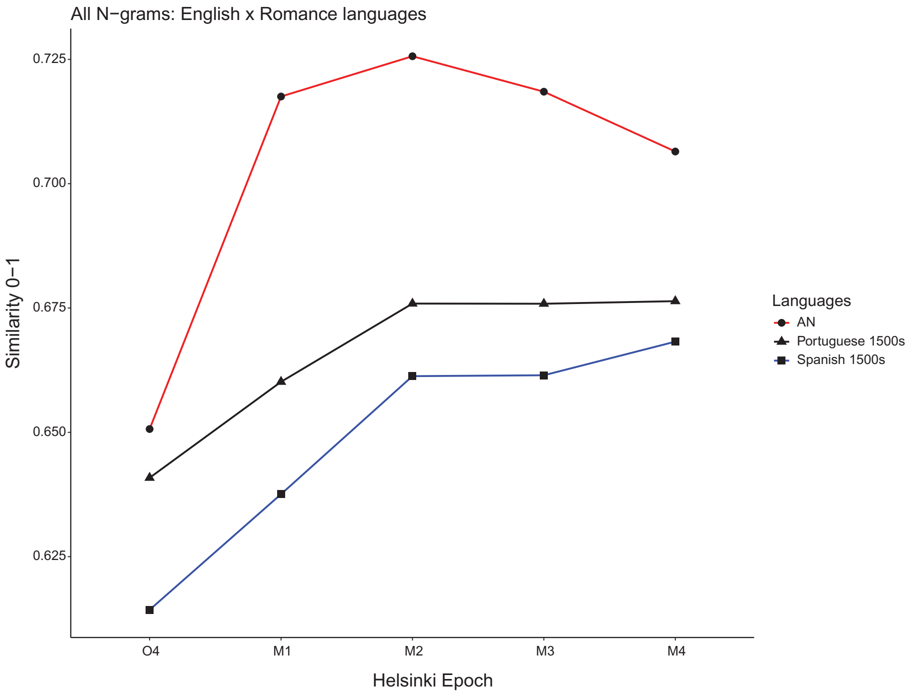
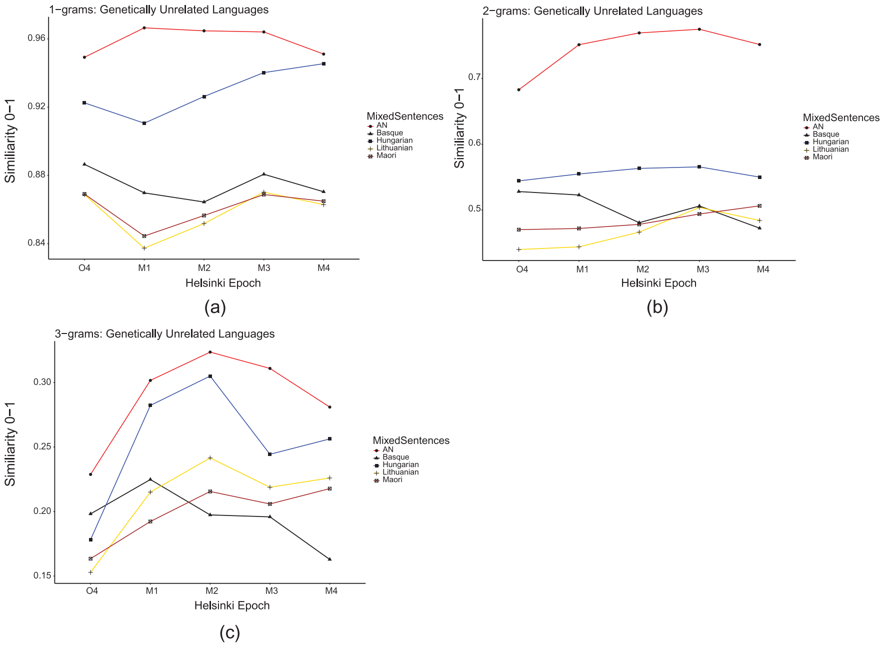


Figure 7: AN and Romance languages similarity with English across time.

languages. This signals caution in analyzing change between languages without adequate comparison languages. It could be errantly concluded that English had contact with Portuguese if viewed in isolation without comparison to French also.

3.10 Unrelated or genetically distant language comparisons

Given that Basque demonstrated a fluctuating similarity of uncertain origin, a comparison of English over time to three further unrelated languages was performed (see Section 2.1.2). This was to further examine the level of chance resemblance and its variance across time with English. Hungarian, Lithuanian, and Maori's N-gram profiles were compared to English. Figure 8 charts these similarity trends over time. Surprisingly all non-contact languages examined showed to some degree the same relative pattern.



(a)
(b)
(c)

Figure 8: Non-contact languages' similarity with English over time.

Across the 4 non-contact languages, Unigrams similarity shows a similar trend with a decrease in similarity between O4 and M1 and a subsequent increase in similarity between M1 and M3. Similarly for bigram profiles, there is an increase in similarity at M1 to M3 while there is no downward trend between O4 and M1. Trigrams on the other hand show the opposite pattern with a sharp increase in similarity after O4 and decreases in similarity.

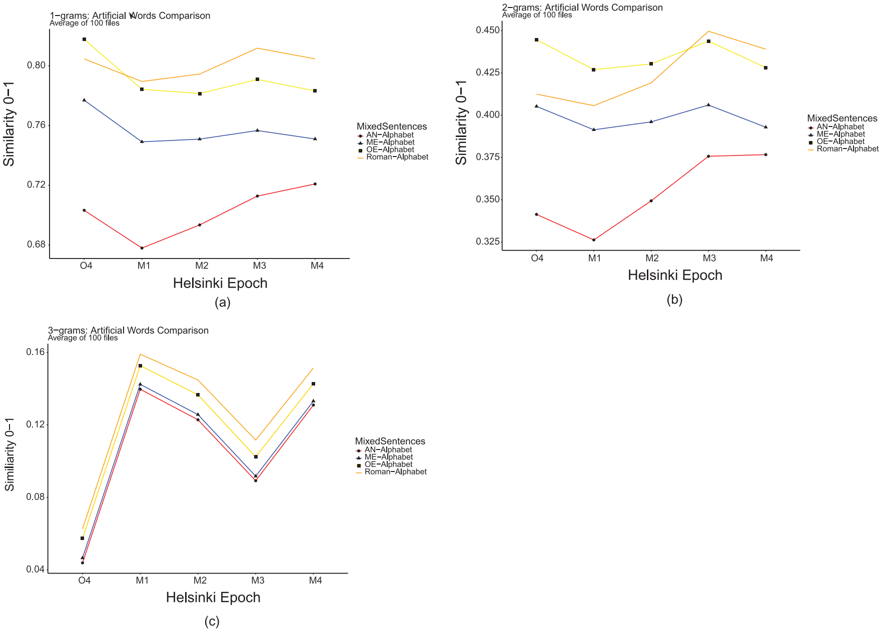
It was sought to rule out any affect of shared vocabulary. It was decided to create samples of artificial text written in several alphabets. 4 alphabets were chosen. Firstly the 26 letter Roman alphabet. Secondly the alphabet used by OE including thorn and eth, but excluding <z> and <v>, in order to see if any of these changes in similarity were responding to the fluctuations of Anglo-Norman orthography. Furthermore, the alphabet used by the Anglo-Norman texts was used and lastly the alphabet used in all ME periods was used, including letters not used in OE such as yogh, <’>, and <~>.

Several syllable structures were defined (such as “CVC”; see Supplementary Materials). Each word was randomly assigned a syllable count between 1 and 5. A structure for each syllable was randomly chosen. Consonants and vowels were randomly supplanted into the structure to form artificial words. A dictionary of 5000 artificial words was created per alphabet. To create each sample of artificial text, a word was selected at random from the dictionary list. Each file consisted of 10,000 words. 100 files for each alphabet were created. The similarity values displayed an average of these 100 files.

Figure 9 surprisingly displays a similar pattern shown by the real-world languages. Unigrams show a decrease between O4 and M1 and an increase between M1/2 and M3. Bigrams show the upward trend at M2/3 also. However the artificial texts show a downward trend between O4 and M1 that the real world files did not display. The pattern for trigrams mirrors the real-world languages with an increase in similarity between O4 and M1. This would suggest that it is not the content of the languages creating this fluctuation.

A Chi Square analyses were performed between two-selected non-contact languages, Hungarian and Basque, and the epochs of English, in an effort to detect the cause of these fluctuations. The divergence of N-grams from English in each period was cross-compared and it was investigated as to which English N-grams the unrelated languages were growing more close or distant to.

Table 10 shows the changes in the divergence in features of Basque and Hungarian to English between O4 and M1. The highlighted unigrams indicate that the downward trend is influenced by M1’s new orthographic elements, such as yogh and crossed thorn, which were shown to have increased in frequency in Table 4. The OE vector did not contain these features thus when they are introduced the number of features not shared by the non-contact languages increases, furthering the distance between the languages. The increase in similarity found between trigram profiles seems to be driven by features that were highlighted in Table 4 as having decreased in frequency between O4 and M1,



(a)

(b)

(c)

Figure 9: Artificial text samples' similarity to English over time.

such <þæt> and <þam>. Thus conversely as English loses N-grams between periods, the unrelated languages register as closer.

Table 11 displays the increases in unigrams and bigrams that possibly drive the increase in similarity between M2 and M3. The features found are those found to have begun to decrease in Late ME in Table 6. As Early ME thorn and yogh drop from usage, the unrelated languages measure as more similar. Thus this waxing and waning similarity in non-contact languages is an impact of the changes of within English between periods. Future research might apply some form of baseline correction. The similarity increases between French and Latin might be artificially boosted by these fluctuations or measured as less than it should be.

3.11 Change in English affixes

In order to further elucidate what specifically Anglo-Norman influenced within English, prefix and suffix N-grams were generated. For example a bigram suffix was two slices from the end of the word, excluding the white space. The affixes of each period was compared to the sample of Anglo-Norman across the ME period.

Table 10: Unigram and Trigram changes between O4 > M1 and Basque/Hungarian.

Basque				Hungarian			
N-gram	***Residual O4***	***Residual M1***	***Diff.***	***N-gram***	***Residual O4***	***Residual M1***	***Diff.***
1-grams							
ʒ	0	−62.3	*−62.3*	**ʒ**	0	−21.6	*−21.6*
þ*	0	−41.9	*−41.9*	**e**	−19.3	−35.9	*−16.5*
þ	−94.8	−118.4	*−23.5*	**ꝥ***	0	−14.6	*−14.6*
ð	−74.1	−95.3	*−21.2*	**h**	−2.6	−17.1	*−14.5*
w	−40.8	−61.2	*−20.4*	**u**	−24.0	−33.4	*−9.4*
3-grams							
þæt	−38.2	−5.6	*32.6*	**_ge**	−24.0	−3.8	*20.2*
_þæ	−48.2	−18.3	*29.8*	**þæt**	−16.2	−1.9	*14.3*
æt_	−42.8	−14.1	*28.6*	**_þæ**	−20.	−6.3	*14.1*
um_	−26.0	−4.8	*21.2*	**æt_**	−18.2	−4.9	*13.2*
cyn	−21.7	−4.8	*16.8*	**um_**	−17.1	−5.6	*11.5*
_cy	−23.8	−8.2	*15.6*	**on_**	−21.5	−11.1	*10.3*
gew	−19.7	−4.7	*14.9*	**ges**	−9.3	−0.7	*8.6*
þam	−26.9	−12.6	*14.2*	**cyn**	−9.2	−1.6	*7.5*
yng	−18.9	−4.8	*14.1*	**hi_**	−10.2	−2.9	*7.2*
ig_	−18.6	−6.5	*12.1*	**_cy**	−10.1	−2.8	*7.2*
ges	−26.8	−14.9	*11.8*	**þam**	−11.4	−4.3	*7.0*
m_g	−19.6	−7.8	*11.7*	**eal**	−15.3	−8.3	*6.9*
æs_	−29.8	−18.6	*11.2*	**dan**	−8.4	−1.4	*6.9*
wæs	−24.3	−13.1	*11.1*	**gew**	−8.4	−1.6	*6.7*
að_	−24.0	−12.9	*11.0*	**_ea**	−16.0	−9.4	*6.6*

In Figure 10 suffix slices are shown to grow more similar to French throughout ME, while prefix slices remain relatively uninfluenced until later periods. For 1-character slices, prefix slices remain low in similarity until M3. 2/3-character slices correspond more to the sharp increase in French similarity shown in Figure 6. By the end of ME, suffixes still show a greater similarity to prefixes, indicating that suffixes were the carrier of some of the French influenced change.

3.12 French influence on the most frequent vocabulary

In order to examine the effect of French on English vocabulary, the corpus was split by word usage frequency. Three frequency bands were created, the first hundred most frequent words, followed by the second and third hundred in each

Table 11: Unigram and Bigram changes between M2 > M3 and Basque/Hungarian.

Basque				Hungarian			
N-gram	***Residual M2***	***Residual M3***	***Diff.***	***N-gram***	***Residual M2***	***Residual M3***	***Diff.***
1-grams							
~	0	9.8	*9.8*	**þ**	−46.9	−35.8	*11.0*
ʒ	0	2.7	*2.7*	**e**	−32.8	−26.1	*6.7*
				ʒ	−21.3	−15.2	*6.1*
				‘	0	4.7	*4.7*
				u	−33.1	−28.7	*4.3*
2-grams							
_ð	−20.3	−7.6	*12.7*	**þe**	−32.4	−23.0	*9.3*
iʒ	−30.4	−20.1	*10.2*	**_þ**	−40.2	−32.7	*7.5*
eþ	−34.4	−24.7	*9.7*	**þ_**	−10.8	−17.8	*6.9*
uo	−25.7	−16.0	*9.6*	**ne**	−7.0	−0.7	*6.2*
ʒt	−37.4	−28.0	*9.3*	**de**	−24.7	−18.4	*6.2*
þ_	−46.8	−37.9	*8.9*	**im**	−11.5	−5.2	*6.2*
‘_	−15.6	−7.3	*8.2*	**ʒt**	−14.2	−7.9	*6.2*
yþ	−16.4	−8.5	*7.8*	**eþ**	−13.1	−7.0	*6.0*
j_	−18.9	−11.3	*7.5*	**iʒ**	−11.5	−5.7	*5.8*
þo	−37.2	−29.8	*7.4*	**þo**	−14.1	−8.5	*5.6*

epoch (see Supplementary Materials for sample of these frequency bands). The frequency of each word in the period was calculated and each word was assigned to a frequency bin. Character N-gram profiles were then generated for each subdivision of the period. This had a weakness in that it would have removed spelling variants which would be lower occurring than their dominant spelling and thus would have likely not been frequent enough to be included in each of the three frequency bands.

Firstly internal similarity across time was calculated for each frequency band. O1 was excluded due to its extremely low relative word count. Figure 11 displays the similarity across time for each band. The top frequency band changes the least across time indicating that its contents remain similar across time. The top 100 most occurring items could be considered to be the core vocabulary of the language. Thus this lends further empirical proof to the concept of a core basic vocabulary that changes little over time. The second and third hundred most frequent words decrease in similarity at roughly the same rate. This provides some indication that the cause of their variance may be the same.

The similarity between Anglo-Norman and each frequency band was calculated. In Figure 12 the trend of similarity can be seen. Interestingly the top 100

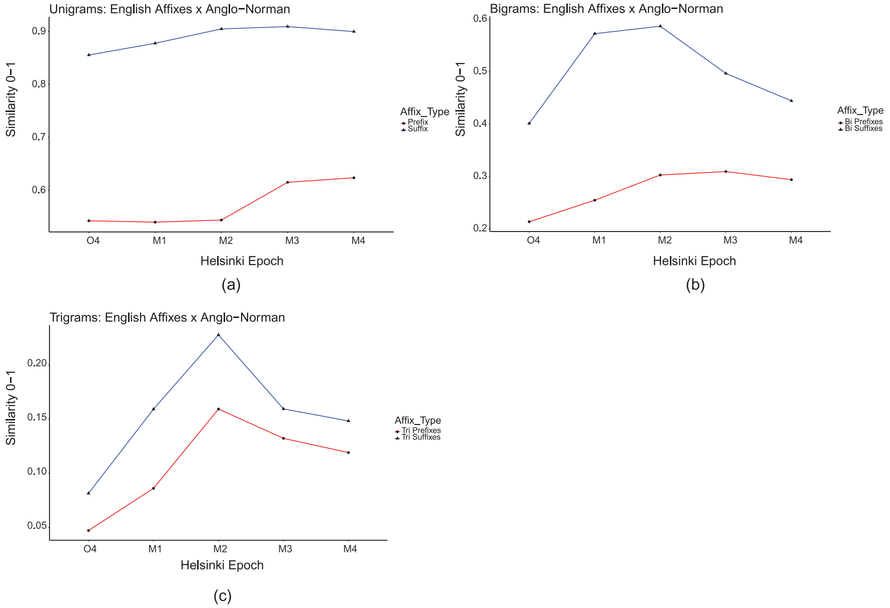


Figure 10: Affixes similarity between English and Anglo-Norman over time.

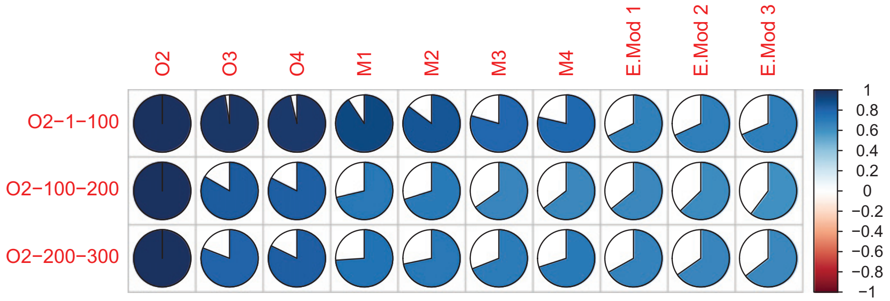


Figure 11: English word-frequency bands internal diachronic change.

most occurring items for each period is the least similar to French across the entire period of contact, showing only minor increases in similarity. This lends empirical evidence to the concept of the core vocabulary that is resistant to borrowing. However some increase in similarity to French is detectable in the top frequency band, possibly indicating how the Swadesh-200 list of core items was able to detect an increase in French similarity.

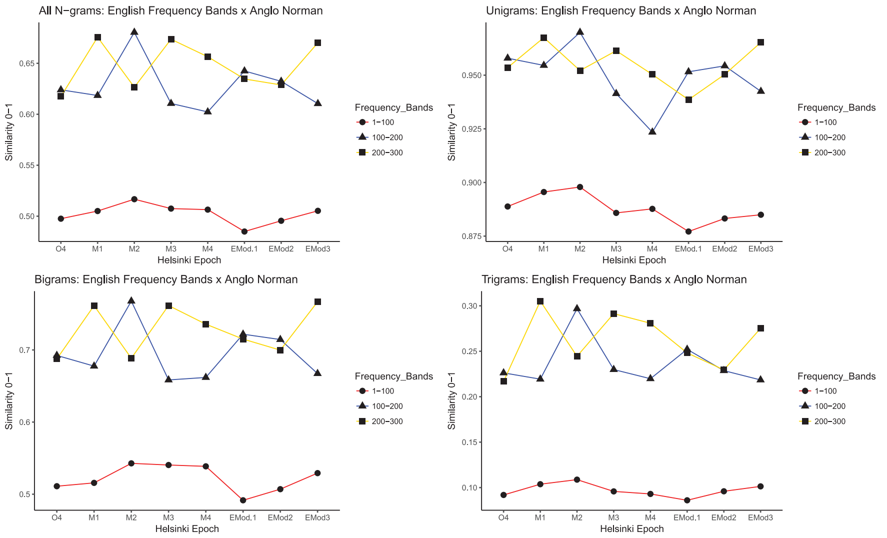


Figure 12: English word-frequency bands x Anglo-Norman over time.

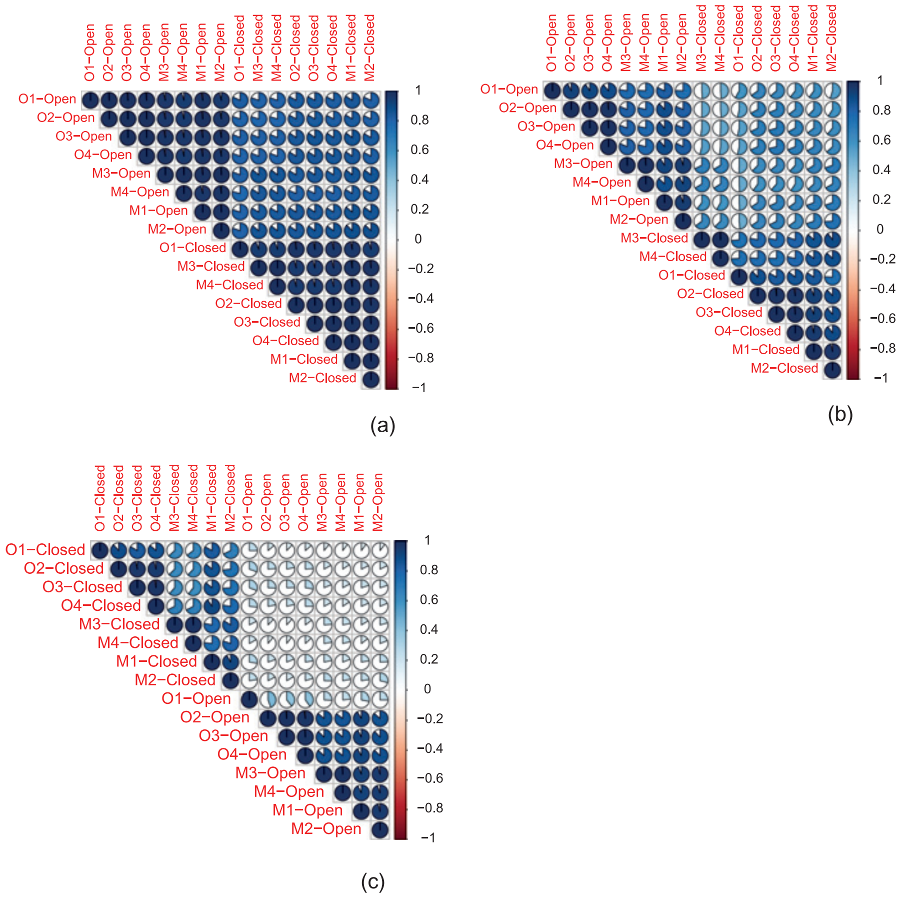
The second and third frequency bands show a pattern roughly contrary to one another, with one showing a rise and the other a decrease. The third hundred most frequent items show the greatest influence of French at first, before at M2, the second frequency band shows a sharp rise. This possibly reflects the degree to which French vocabulary affected the most commonly expressed concepts. M2 being a period of high French influence, French appears to have influenced the most frequent non-basic lexical items (the 2nd frequency band) before waning in influence by M3. While in M3, the 3rd most frequency items display a similar growth in French influence. With a finer temporal resolution, future studies could examine the exact time course of this growth in influence.

3.13 French Influence on grammatical items

The extent to which this top frequency band was composed of grammatical items was unknown. It was sought to examine whether similar results could be generated from an analysis of grammatical as compared to non-grammatical items, following from the distinction between closed class and open class items.

A top-down filter was applied to the data that separated grammatical words from the text. A dictionary of the most prominent grammatical words was compiled. For OE, each grammatical word, including determiners, pronouns, numbers, conjunctions etc., were taken from Mitchell and Robinson (2011) and Baker (2012). Given the orthographic variation around <þ> and <ð>, multiple versions of each word were created with the variant. For the ME period, grammatical items from Burrow and Turville-Petre (2013) were collated. For the transitional periods in Early ME, a combined dictionary of OE and ME items were applied. Each word on the dictionary list was removed from the text file and collected in a new sample text. Due to spelling variations some closed class words will have been missed and still appear in the open class files. N-gram profiles were generated for each sample.

Firstly the internal similarity across time was computed. Hierarchical clustering through the R package corrplot (Wei and Simko 2013) was applied to examine which items remained most similar across time. In Figure 13, across all N-grams



(a)

(b)

(c)

Figure 13: English open x closed class internal diachronic correlograms.

open class items remain most similar to open class items and closed class to closed class across time despite the internal change across time in both categories.

This provides validation that data-driven methods can derive similar results to analyses that apply top-down filters such as the grammatical item list above. It furthers provides support that the most frequent items in each period were likely to be partly grammatical words due to the similar patterns displayed.

Similarity of open and closed class categories to French were calculated. In Figure 14, open classes show the greatest similarity to French and the greatest increase post-conquest. Unigrams show an increase in French similarity in closed classes, possibly indicating that grammatical items were influenced in their orthography while not in their content as shown by trigram similarity that remains consistently low. This brings top-down verification that closed class grammatical items are less effected by borrowing than open classes.

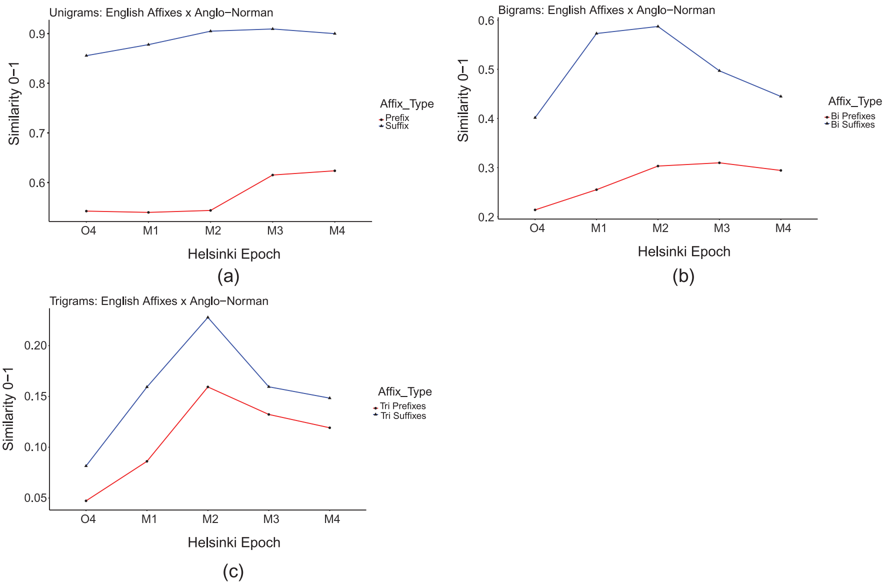


Figure 14: English open and closed class x Anglo Norman.

3.14 Examination of divergence within a language family

Lexicostatistical studies are often interested in determining genetic relations among languages and reconstructing language families. The diversification of OE from its nearest linguistic neighbour Old Frisian was measured using N-gram profiles.

In Figure 15, the similarity of Frisian to English across time was calculated. A surprising pattern was measured. It was expected that a decrease in similarity would be measured, as English diversifies away from its Germanic relative towards the Romance language French. However Old Frisian paradoxically shows an increase in similarity in the ME period.

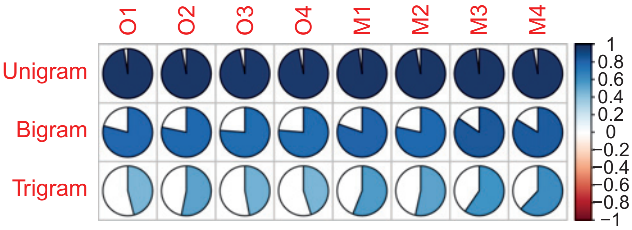


Figure 15: Frisian x English over time.

It is important to remember this analysis of text corpora uses a measure of orthographic similarity not phonetic. OE and Frisian are genetically related and not necessarily orthographically similar. Compare OE <reht> to Frisian <riuht> ‘law’ or OE <cirice> to Frisian <tsyurka> ‘church’ (Robinson 2005). Each item is phonetically similar yet orthographically distant. While labeled “Old” it is in fact concurrent with most “Middle” stages of medieval Languages (Robinson 2005). With this in mind, the increase in similarity to ME, might reflect a similarity to temporally contiguous periods. With similar orthographic representations, such as <th> and <k> there may be some ground for further studies of co-varying orthographies across time.

Features of difference were examined using Chi Square. In Table 12, A representative period of OE was chosen (O3) and its feature’s frequency counts were compared to Frisian. It is apparent that Frisian is indeed orthographically distant from OE. Frisian is lacking OE ubiquitous characters such as thorn, eth, and ash. Frisian lacks OE’s representation of diphthongs with a large difference in <ea> and <eo>. Indications of a morphosyntactic difference can be seen in the lack of the prefix <_ge> and the verbal infinitive marker <an_>.

While Frisian, on the other hand, uses orthography seen in English in ME, thus explaining the rise in similarity. Common ME letters such as <k>, <th>, and French origin <ch> feature, along with letter combinations such as <ent> which overlaps with French imported suffix into ME <ent> (see Table 10).

Table 13 demonstrates the increase in similarity of Frisian to English between OE and ME calculated through a word-sense list. Both a phonetic and

Table 12: Frisian x OE (O3) – Contrasted features (derived via X^2).

N-Gram	*Residual*	*P value*	X^2	*N-Gram*	*Residual*	*P value*	X^2
1gram				1gram			
Positive				Negative			
Difference				Difference			
t	**107.8**	*0.00*	*18,708*	æ	**−108.6**	*0.00*	*18,382*
k	**94.1**	*0.00*	*13,630*	þ	**−106.1**	*0.00*	*17,533*
h	**69.6**	*0.00*	*7723*	ð	**−96.7**	*0.00*	*14,528*
j	**68.7**	*0.00*	*7743*	g	**−68.2**	*0.00*	*7323*
i	**54.7**	*0.00*	*4834*	o	**−53.6**	*0.00*	*4622*
2gram				2gram			
Positive				Negative			
Difference				Difference			
th	**154.6**	*0.00*	*37,030*	_þ	**−93.3**	*0.00*	*13,520*
_t	**114.3**	*0.00*	*20,269*	ea	**−73.4**	*0.00*	*8322*
en	**85**	*0.00*	*11,258*	eo	**−70.8**	*0.00*	*7751*
er	**79.5**	*0.00*	*9806*	ge	**−65.2**	*0.00*	*6600*
i_	**75.3**	*0.00*	*8736*	_g	**−63.3**	*0.00*	*6207*
ch	**75**	*0.00*	*8659*	ð_	**−55.3**	*0.00*	*4708*
et	**68.8**	*0.00*	*7309*	þæ	**−55.2**	*0.00*	*4698*
ha	**68.2**	*0.00*	*7178*	_ð	**−54.6**	*0.00*	*4595*
a_	**66.1**	*0.00*	*6833*	æt	**−53.5**	*0.00*	*4407*
ke	**61.6**	*0.00*	*5822*	þa	**−47.1**	*0.00*	*3549*
nt	**58.4**	*0.00*	*5348*	þe	**−46**	*0.00*	*3410*
3gram				3gram			
Positive				Negative			
Difference				Difference			
_th	**181.8**	*0.00*	*27,210*	_ge	**−63.37**	*0.00*	*6193*
the	**91.6**	*0.00*	*12,917*	_þæ	**−54.2**	*0.00*	*4525*
tha	**91.6**	*0.00*	*8719*	æt_	**−47.8**	*0.00*	*4869*
ha_	**72.2**	*0.00*	*8002*	e_þ	**−46.6**	*0.00*	*3338*
_en	**69.9**	*0.00*	*7504*	on_	**−46**	*0.00*	*3272*
thi	**68.9**	*0.00*	*7292*	þæt	**−43.8**	*0.00*	*2955*
et_	**67.4**	*0.00*	*6984*	_þa	**−43.6**	*0.00*	*2928*
cht	**58.7**	*0.00*	*5300*	an_	**−41.8**	*0.00*	*2707*
ent	**57.8**	*0.00*	*5129*	_on	**−41.4**	*0.00*	*2646*
th_	**57**	*0.00*	*4982*	eor	**−39.4**	*0.00*	*2391*
nte	**56.2**	*0.00*	*4857*	eal	**−39.3**	*0.00*	*2377*
er_	**56.1**	*0.00*	*4845*	_ea	**−38**	*0.00*	*2227*
her	**55.7**	*0.00*	*4776*	þe_	**−36.1**	*0.00*	*2009*
sen	**52.1**	*0.00*	*4166*	þa_	**−34.9**	*0.00*	*1987*
het	**51.4**	*0.00*	*4057*	æs_	**−34.7**	*0.00*	*1849*

Table 13: Change in Frisian similarity between OE and ME.

	Orthographic	Phonetic
Difference in %	0.017	0.007

orthographic transcription display little change between the periods. The amount of words used was perhaps not great enough to show the expected difference due to differing orthography. Orthography displays only a 1% increase in similarity. More surprisingly however is the lack of any decrease between the periods. While the Swadesh lists may reveal a closeness of English to its nearest neighbour Frisian, it may not be sensitive enough to detect any finer differentiation between periods of English, at least for this small sample.
A comparison of an orthographic and a phonetic transcription was attempted for the text corpora samples. The corpora have the advantage of incorporating not only the core lexicon but also a wider vocabulary and the morphology of a language. A subsection of the Frisian Corpus, the Codex Unia, was selected. One text was chosen to hopefully decrease the variation in spelling and the Latin therein was successfully removed from this sample (see Section 2.1.2). Helsinki Epochs O3 and M4 were chosen as they are late stages in their respective epochs and their orthographic variation may be less than earlier periods of greater change, allowing for a more accurate phonetic transliteration.

Figure 16 demonstrates the increase in orthographic similarity found previously by both the N-gram based method and the lexicostatistical method. However once the text corpora are transliterated into a common phonetic transcription (the ASJP code), the expected decrease in similarity can be seen. This suggests that N-gram profiles, provided a phonetic transcription, can detect divergence between languages where a Swadesh list does not.

4 Discussion

The analyses above have for the most part validated the use of character N-grams as a means to quantify and explore historical language variation. The measures were sensitive enough to derive the majority of the outcomes expected from the qualitative literature. Briefly, distance measures between frequency profiles of N-gram features were able to derive the expected decrease in internal similarity between Early OE and Late ME. Between languages, a quantification of the period of language contact between English and

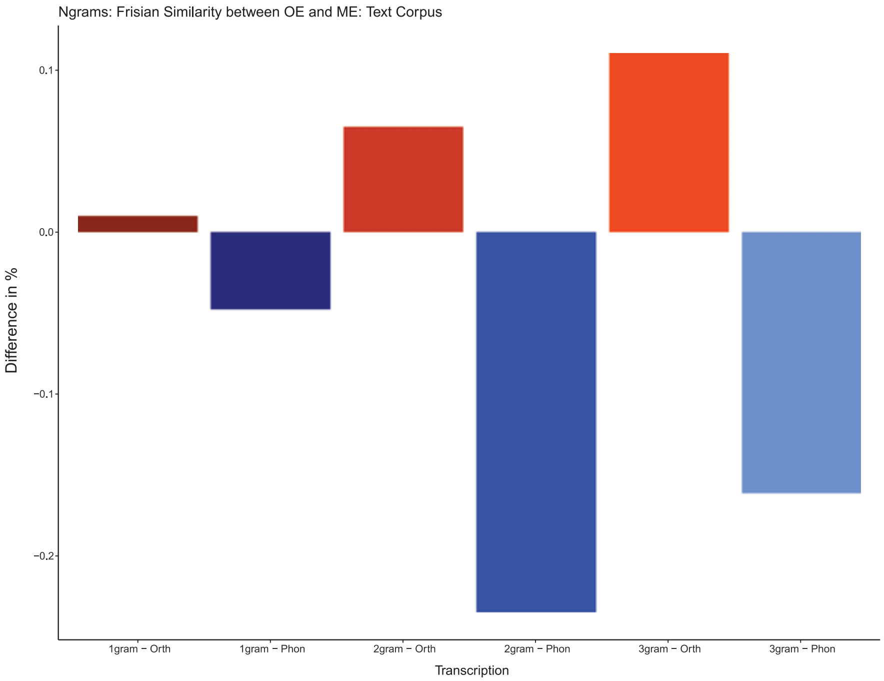


Figure 16: Change in Frisian similarity between OE and ME – Phonetic vs orthographic.

French was calculated. The similarity metric was successful in highlighting French over Latin as the Romance language of greatest influence. However the method was inconclusive in deriving a differentiation between English and Frisian without a transliteration into a common phonetic alphabet. Similarly the method failed to distinguish between influence Anglo-Norman and Continental French on English. It is important to note that the hypotheses outlined in Section 1.5 were fairly broad as the study was largely exploratory and thus less easily falsified.

Similarly an analysis using lexicostatistical methods too detected a decrease in similarity between OE and ME. However when applying this methodology to newly delineated time periods within a gross epoch (in this case Early ME with the P-LAEME corpus), the use of a Swadesh list demonstrated the frailty of the method. No full Swadesh list could be completed and a larger vocabulary list was used instead. This was also seen in the case of Old Frisian where Novotná and Blažek (2007) could neither satisfy a full Swadesh-200 list. While what is an acceptable word count for this type of analysis varies in opinion, there was ample text data in the P-LAEME files compared to other historical languages and the lack of ability to fulfill a list of core words within these sub-periods shows a lack of flexibility to analyses using a finer temporal resolution and by logical extension to spatial resolution.

Most surprising was the lexicostatistical method's ability to detect an influence from French, primarily orthographically. This highlights the importance of the choice transcription. Orthographic variation in English is an area of change itself and not an obscurant to phonetic change. Thus the goal of the study must guide the top-down augmentation to a phonetic alphabet in that if genetic inheritance is desired as in lexicostatistical reconstruction of family trees, a phonetic script might reveal this better as in the case of Frisian (see Section 3.4).

The patterns detected with lexicostatistical methods were also found with the less expert dependent N-gram method. The use of N-gram frequency profiles proved more flexible in its application to many periods of English, even periods of lower word counts where the construction of a Swadesh list would be less likely to be successful such as period O1. In terms of examining contact, the lexicostatistical method only provided a rough picture, merely confirming that there was change. While the comparison of character N-gram profiles between languages detected a richer pattern that corresponded well to the highs and lows of Romance vocabulary influence identified through the literature (e.g. Durkin 2014). It is important to note however that as there is no quantitative trend to confirm these patterns against it is hard to measure what was not detected.

Automatic cognate detection has been put to use to automatically create word-sense lists. Scherrer (2012) applied a cognate detection algorithm to highlight cognates between texts of Swiss German dialects. A normalized Levenshtein distance (see Section 2.1.3) was used to measure the distance of words between the text and words with a distance above a threshold were deemed as cognates. Through this, Scherrer was able to bypass the laborious collation of a hand-drawn list for each dialect region, suggesting this could also be applied to historical epoch and historical dialects synchronically and diachronically. However automatic cognate detection based on word-to-word distances might be have difficulty in picking up extreme variants such as ME <trghug> and <yruȝ>, "through", rendering a LDN distance of 0.33, below Scherrer's most successful thresholds. Thus these methods might only detect less variable cognates. Scherrer's evaluation relied on confirming as correct cognate pairs detected. Yet the true number of cognate pairs missed is not assessed. High thresholds would correctly identify *some* cognate pairs but leaving an unknown amount unidentified possibly due to more extreme variation. Automatic cognate detection methods should be tested on datasets of confirmed cognates in a chosen text so that the accuracy rates in detecting all cognates could be assessed. A similar method that could detect cognates between texts would be helpful in charting loanwords from French into English. However this would solely be an analysis of lexicon and be less able to comment on orthographic or morphological exchange between languages as described in Section 1.4.

A further avenue of research could harness the orthographic and phonetic variation within historical English towards a practical goal. Wahlberg et al. (2016), Zampieri et al. (2016) and Szymanski and Lynch (2015) all used computational methods involving word or character N-grams in diachonric corpora towards temporal text classification. Character N-gram profiles of historical English could be used to classify a text of unknown date to a historical period. This would be of great benefit to researchers in and outside of Linguistics as placing a text in its proper historical epoch is necessary to understand the text fully. By logical extension this could be applied to other variables such as geographic location, in other words detecting the dialect of a text. Faulkner (2017) in examining overlooked English glosses qualitatively compared their spellings in order to place them to some temporal period and some dialect. For instance, features in the glosses such as <_hr> were used by him to place the words in the OE period rather than the ME period. The current method could provide datasets of varying features over time and geography for researchers to compare such data to. Using methods such as Chi Square positive and negative differences between periods, spatial regions, and spatial regions over time could be measure and in effect create bottom-up dialect and temporal maps.

Finally French influence can be examined across many text groupings, time periods, geographic regions and text types, allowing for a quantitative database of French influence in England to be compiled. Further research could also trial the ability to detect contact in other settings historical and modern.

Acknowledgements: This research is supported in part by the Science Foundation Ireland (grants 12/CE/I2267 and 13/RC/2106) through the CNGL programme and the ADAPT Centre.

References

Alcorn, Rhona, Robert Truswell, Joel Wallenberg & James Donaldson. 2018. A parsed linguistic atlas of early middle English. https://datashare.is.ed.ac.uk/handle/10283/3032 (accessed April 2018).

Baker, Peter S. 2012. *Introduction to old English*, 3rd edn. Oxford: Wiley-Blackwell.

Benoit, Kenneth & Paul Nulty. 2013. Quanteda: Quantitative analysis of Textual Data, An R library for managing and analyzing text.

Borin, Lars. 2013. The why and how of measuring linguistic differences. *Approaches to measuring linguistic differences*, 3–25. Berlin: Walter de Gruyter.

Borin, Lars, Bernard Comrie & Anju Saxena. 2013. The intercontinental dictionary series—A rich and principled database for language comparison. *Approaches to measuring linguistic differences*, 285–302. Berlin: Walter de Gruyter.

Bremmer Jr, Rolf H. 2009. *An introduction to Old Frisian: History, grammar, reader, glossary*. Amsterdam: John Benjamins Publishing.

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

Brown, Cecil H., Eric W. Holman, Søren Wichmann & Viveka Velupilla. 2008. Automated classification of the world's languages: A description of the method and preliminary results. *STUF-Language Typology and Universals Sprachtypologie Und Universalienforschung* 61(4). 285–308.

Buchta, Christian, Kurt Hornik, Ingo Feinerer & David Meyer. 2017. Package 'tau'.

Burrow, John Anthony & Thorlac Turville-Petre. 2013. *A book of Middle English*. Oxford: John Wiley & Sons.

Cavnar, William B. & John M. Trenkle. 1994. N-gram-based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US. 161–175

Chadwyck-Healey. 1995. Patrologia Latina the full text database. Alexandria, VA: Chadwyck-Healey. http://pld.chadwyck.co.uk/ (accessed September 2017).

Cinque, Guglielmo. 2006. *Restructuring and functional heads: The cartography of syntactic structures volume 4: The cartography of syntactic structures*, vol. 4. New York, USA: Oxford University Press.

CLUL (ed.). 2014. *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. http://ps.clul.ul.pt (accessed April 2018).

Dalton-Puffer, Christiane. 1996. *The French influence on Middle English morphology: A corpus-based study on derivation*. Berlin: Walter de Gruyter.

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

Darwin, Jason & Māmari Stephens. 2009. The legal Māori archive: Construction of a large digital collection. dx.doi.org/10.2139/ssrn.2548118, http://nzetc.victoria.ac.nz/tm/scholarly/tei-legalMaoriCorpus.html (accessed April 2019).

de Saint-Exupéry, Antoine. 1948. *Le Petit Prince*. Paris, France: Éditions Gallimard.

Durkin, Philip. 2014. *Borrowed words: A history of loanwords in English*. Oxford: Oxford University Press.

Einhorn, Einar. 1974. *Old French: A concise handbook*. Cambridge: Cambridge University Press.

Ellison, T. Mark & Luisa Miceli. 2012. Distinguishing contact-induced change from language drift in genetically related languages. Proceedings of the Workshop on Computational Models of Language Acquisition and Loss, Association for Computational Linguistics.

Etchegoyhen, Thierry, Andoni Azpeitia & Naiara Pérez. 2016. Exploiting a large strongly comparable corpus. Proceedings of the 10th edition of the Language Resources and Evaluation Conference, European Language Resources Association (ELRA), Paris, France.

Faulkner, Mark. 2017. Dublin, trinity college, MS 492: A new witness to the Old English Bede and its twelfth-century context. *Anglia* 135(2). 274–290.

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

Freeborn, Dennis. 2006. *From Old English to Standard English*, 3rd edn. Basingstoke: Macmillan.

Fuka, Karel & Rudolf Hanka. 2001. Feature set reduction for document classification problems. *IJCAI-01 Workshop: Text Learning: Beyond Supervision*.

Gudschinsky, Sarah C. 1956. The ABC's of lexicostatistics (glottochronology). *Word* 12(2). 175–210.

Guillot, Céline, Alexei Lavrentiev & Christiane Marchello-Nizia. 2007. La Base de Français Médiéval (BFM): états et perspectives. In Pierre Kunstmann & Achim Stein (eds.), *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 Février 2006*. Stuttgart: Steiner.

György, Rónay. 1993. *A kis herceg*. Budapest, Hungary: Móra Ferenc Könyvkiadó.

Hogg, Richard. 1992. Phonology and morphology. In Richard Hogg (eds.), *The Cambridge history of the English language, Vol. I: The beginnings to 1066*, 67–167. Cambridge University Press.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, et al. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52(6). 841–875.

Horobin, Simon & Jeremy J. Smith. 2002. *An introduction to Middle English*. USA: Oxford University Press.

Ingham, Richard. 2011. Anglo-Norman correspondence corpus (Online). http://wse1.webcorp.org.uk/anglo-norman/about.html (accessed November 2017).

Kauneckas, Vytautas. 1998. *Mažasis Princas*. Vilnius, Lithuania: Džiugas.

Key, Mary Ritchie & Bernard Comrie (eds.). 2015. *The intercontinental dictionary series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://ids.clld.org (Accessed May 2017).

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 97–133.

Kuhn, M. 2012. *The caret package*. http://cran.r-project.org/web/packages/caret/caret.pdf.

Kytö, Merja. 1996. *Manual to the diachronic part of the Helsinki Corpus of English texts. Coding conventions and lists of sources*. Helsinki, Finland: University of Helsinki, English Department.

Laing, Margaret & Roger Lass. 2007. *A linguistic atlas of early Middle English, 1150–1325*. http://www.lel.ed.ac.uk/ihd/laeme2/laeme2_framesZ.html (accessed September 2017).

Lass, Roger (ed.). 2000. *The Cambridge history of the English language*, vol. 3. Cambridge, UK: Cambridge University Press.

Lee, Yeon-Ju & Laurent Sagart. 2008. No limits to borrowing: the case of Bai and Chinese. *Diachronica* 25(3). 357–385.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8). 707–710.

Liaw, Andy & Matthew Wiener. 2002. Classification and regression by randomForest. *R News* 2(3). 18–22.

McMahon, April & Warren Maguire. 2011. Quantitative historical dialectology. In D. Denison & R. Bermúdez-Otero (eds.), *Analysing Older English*, 140–158. Cambridge: Cambridge University Press.

Millar, Robert McColl & Robert L. Trask. 2015. *Trask's historical linguistics*. London: Routledge.

Mitchell, Bruce & Fred C. Robinson. 2011. *A guide to Old English*. Chicester: Wiley-Blackwell.

Novotná, Petra & Václav Blažek. 2007. On lexicostatistic classification of the Frisian dialects. *Linguistica Brunensia. Sborník prací filozofické fakulty brněnské university* 56. 115–132.

Piotrowski, Michael. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies* 5(2). 1–157.

Rama, Taraka, G. K. Mikros Lars Borin & J. Macutek. 2015. Comparative evaluation of string similarity measures for automatic language classification. In George K. Mikros & Jan Macutek (eds.), *Sequences in language and text*, 171–200. Berlin\Boston, Germany: De Gruyter Mouton.

Robinson, Orrin W. 2005. *Old English and its closest relatives: a survey of the earliest Germanic languages*, 2nd edn. Stanford, CA: Stanford University Press.

Rothwell, William & D. A. Trotter. 2008. *Anglo-Norman dictionary*. Aberystwyth University and Swansea University. http://www.anglo-norman.net.

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

Scherrer, Yves, 2012. Recovering dialect geography from an unaligned comparable corpus. Proceedings of the EACL 2012 Workshop on Visualization of Language Patterns and Uncovering Language History from Multilingual Resources, Avignon.

Shannon, Thomas. 1999. Corpus of Old East Frisian Texts. http://titus.uni-frankfurt.de/texte/etcs/germ/afries/afrcorp/afrco.htm/ (accessed September 2017)

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

Singh, Anil Kumar & Harshit Surana. 2007. Can corpus based measures be used for comparative study of languages? Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology, Association for Computational Linguistics.

Sytsema, J., J. A. Nijdam, H. D. Meijering, O. Vries & P. Stiles. 2012. A diplomatic edition of Codex Unia. http://tdb.fryske-akademy.eu/tdb/index-unia-en.html# (accessed September 2017).

Szymanski, Terrence & Gerard Lynch. 2015. UCD: Diachronic text classification with character, word, and syntactic n-grams. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).

Tomović, Andrija & Predrag Janičić. 2007. A variant of n-gram based language classification. In Roberto Basili & Maria Teresa Pazienza (eds.), *Congress of the Italian association for artificial intelligence*, vol. 4733. Berlin, Heidelberg: Springer.

Trask, Robert L. 1995. Origins and relatives of the Basque language: Review of the evidence. In Jose Ignacio Hualde, Joseba A. Lakarra & Robert Trask (eds.), *Towards a history of Basque language*, 65–99. Amsterdam: John Benjamins.

Vaamonde, Gael, Rita Marquilhas, Ana Luísa Costa & Clara Pinto. 2014. Post Scriptum: Archivo digital de escritura cotidiana. In Sagrario López Poza & Nieves Pena Sueiro (eds.), *Humanidades Digitales: desafíos, logros y perspectivas de futuro*. [Special issue]. Janus [online] Anexo 1. 473–482.

Versloot, Arjen Pieter & Han Nijdam. 2011. Integrated Frisian language database 2.0. *Language Database*. tdb.fryske-akademy.eu/tdb/index-en.html# (accessed September 2017).

Wahlberg, Fredrik, Lasse Mårtensson & Anders Brun. 2016. Large scale continuous dating of medieval scribes using a combined image and language model. 12th IAPR Workshop on Document Analysis Systems (DAS). 48–53.

Wei, Taiyun & Viliam Simko. 2013. corrplot: Visualization of a correlation matrix. *R Package Version 0.73*.

Wichmann, Søren, Eric W. Holman, Dik Bakker & Cecil H. Brow. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and Its Applications* 389(17). 3632–3639.

Zampieri, Marcos, Shervin Malmasi & Mark Dras. 2016. Modeling language change in historical corpora: the case of Portuguese. Proceedings of Language Resources and Evaluation (LREC). 4098–4104.

Supplemental Material: The online version of this article offers supplementary material. (https://doi.org/10.1515/flih-2019-0012)