sciendo *Studia Anglica Posnaniensia 53 (2018): 129–144*
doi: 10.2478/stap-2018-0006

THE ROUND ALLOGRAPH OF <R> IN LATE MIDDLE ENGLISH[1]

JACOB THAISEN[2]

University of Oslo

ABSTRACT

This paper discusses glyphs of the 2-shaped or "round" allograph of the grapheme <r> with a tag protruding from the lower part of the stem, asking whether their distribution in a corpus of some 600 late Middle English texts can be meaningfully related to these texts' localisation in *A Linguistic Atlas of Late Mediaeval English*. It discusses what localisation expresses, and uses regression modelling to show that there is no co-variation between the texts' paleography and their orthography, although there is a measure of correlation between them. The evidence in favour is that the quantitative analysis identifies localisation in northings as a predictor of the occurrence of the tagged form of the allograph, which occurs at a higher frequency in texts localised below the Midlands line at *c.* 300 northings. The evidence against is the form's scattered distribution according to the localisation variable where co-variation would imply a more clear-cut concentration of points, and also the moderate success at explaining the form's distribution by means of variables known to explain orthographic variation.

Keywords: Middle English; orthography; paleography; regression modeling

1. Introduction

Linguistic levels may blur into each other and largely be predicted by the same variables. The difference between *not* and *nat* as forms of the negation in Middle English is standardly taken to indicate a difference in both orthography

1 The author gratefully acknowledges a residential fellowship at the Netherlands Institute of Advanced Study in the Humanities and Social Sciences. The author thanks the anonymous reviewers for this journal for their helpful comments. The paper is associated with the Middle English Scribal Texts programme at the University of Stavanger.

2 Department of Literature, Area Studies, and European Languages, University of Oslo, P.O. Box 1072, Blindern 0316, Oslo, Norway. jacob.thaisen@ilos.uio.no

and phonology, while those between *him* and *hym* as forms of the object-case masculine personal pronoun and between *it* and *itt* as forms of the nominative-case neuter personal pronoun are both standardly accepted to indicate a difference in orthography only. McIntosh (1974, 1975) labels these two types of difference respectively S (for "spoken") and W (for "written"), and it is variation at these two levels that is the focus of *A Linguistic Atlas of Late Mediaeval English* [LALME] (McIntosh et al. 1986).

A subset of W differences excluded from LALME, however, is the subset of allographic differences. Allographs of the same grapheme differ paleographically rather than orthographically by definition. Thus, the difference between <thou> and <thow> as forms of the second-person personal pronoun is orthographic because "u" and "w" are separate graphemes, whereas that between <þɑt> and <þat> is paleographic since "ɑ" and "a" are separate allographs of a single grapheme. That between <thou> and <thov> is paleographic if "u" and "v" are considered allographs of a single grapheme but orthographic if they are considered separate graphemes. This evident blurring of levels is what makes McIntosh (1974, 1975) propose that allographic variation is not essentially different from orthographic variation, ultimately implying that paleography is a branch of linguistics.

This extension of the object of analysis has long been taking place indirectly through retention of certain allographs in transcription of medieval texts. Editors of medieval texts assign graphemic value to the marks they encounter in them but that value is not universally agreed upon for certain marks and may differ between editions.[3] Pertinent marks include a bar added to the stem of glyphs of the grapheme <h> and a macron placed above two successive minims. If the bar does not serve any linguistic function, the difference between glyphs of the grapheme <h> with and without such a bar is allographic and so paleographic. If the bar represents the grapheme <e>, however, it is interesting at the S level even if it may indicate phonological zero. Similarly, two successive minims with a macron hovering above them may combine to form an allograph of either of the graphemes <n> or <u>, or the macron may in effect itself be an allograph of the grapheme <n> and will be expanded as such in semi-diplomatic transcription.

Other examples do not relate to the set of potential abbreviations but to variation between two alphabetic marks proper. It is in fact not entirely accurate to state that LALME categorically excludes allographs, since the orthographic forms cited in the profiles retain the contrast between "þ" and "y" for the grapheme <þ>, between "z" with and without a bar through its stem for the grapheme <z>, between Carolingian "g" and insular "g" for the grapheme <g>,

[3] Robinson & Solopova (1993) discuss levels of transcription and argue for retention of certain graphetic elements in what is otherwise graphemic transcription.

and between "u" and "v" in transcription of orthographic forms where the editors have deemed it to be of interest to do so.

While not levelling allographs to a single grapheme in transcription is common and facilitates mapping out their occurrence and studying them by the same means as other S or W differences, it is rare to come across published studies by linguists which actually give heed to McIntosh's proposal and pursue this possibility. Paleographers too note that few paleographic forms have been systematically collected from texts written in English, especially from documents (cf. Doyle 1994: 93–94) with a view to such a study. The endeavour has, as far as I have been able to ascertain, been undertaken for no more than a single allographic contrast, that between "þ" and "y" as representations of the grapheme <þ>.[4] This variation appears from visual analysis of distributional maps to be predicted by LALME localisation (Benskin 1982).

2. Methodology

There are at least two meaningful ways of probing whether orthography and paleography co-vary based on LALME and on the corpus from which the present data were extracted. One is to use paleographic profiles (allograph/glyph inventories) to establish an altogether fresh set of localisations for the texts that make up the corpus and subsequently study their possible fit with the LALME localisations, for it is a subset of the LALME corpus. The other is less labour-intensive. If there is co-variation, the allographs will be meaningfully distributed on the plane representing the texts' LALME localisations, since they will be distributed just the same way as the orthographic forms. Benskin (1982) did not produce any fresh localisations, having opted instead for plotting out the distribution of the "y" allograph for the grapheme <þ> on the existing plane. The present study has adopted the same methodology.

Its data relate to allographs of the grapheme <r>, specifically the tagged and untagged forms of the round "r" allograph, which are not conventionally retained in transcription. I extracted them from a subset of the texts which supplied the training data for LALME, as has been mentioned. The subset comprises 604 texts, amounting to the Middle English Grammar Corpus (MEG-C), version 2011.1 (Stenroos et al. 2011),[5] and several further texts never transcribed for that

---

4 Laing (1989), Varila (2014), and Blake & Thaisen (2004) discuss allographic variation in relation to textual studies. Thaisen (2011) discusses abbreviations and subsidiary elements in relation to constraints of time and space in the production of manuscripts.

5 Transcriptions and a catalogue of sources are downloadable from the University of Stavanger at https://www.uis.no/mest. There are slight differences between MEG-C and LALME relating to texts written in highly similar orthography by two or more scribes. MEG-C separates such scribes, whereas LALME conflates them.

version of that corpus. In order to enable me to survey the presence/absence of the two forms of the allograph in this subset, I obtained a photographic reproduction of at least one page from every text thanks to the Middle English Scribal Texts programme.

LALME harvested orthographic forms from shorter texts in their entirety, or in the case of longer texts, from extracts taken from their beginning, middle, and end. The extracts were accepted as being representative of the longer text as a whole only if their respective inventories of orthographic forms exhibited sufficient similarity, for several variables select such forms. How, then, can one be confident that a single, random page taken from among the pages examined by LALME will contain allographs of <r> that adequately sample the text from which the page comes?

The question of whether a single page constitutes an adequate sample requires a longer answer. First, what was collected for LALME was an open set of orthographic forms, whereas what I registered was the presence/absence of a small and closed, predefined set of forms of a single allograph in three positions in the word. There is a mathematical relationship between the length of a text and the likelihood of encountering one of these forms in a specific position such that above a certain threshold, any increase in the amount of text considered will result in an infinitesimal increase in the likelihood of encountering it. While I have not calculated the exact threshold amount of text, it is my experience that a single page will typically contain not only the tagged and untagged forms of round <r> but also most other allographs of <r>, insofar as they are attested at all anywhere in the text.[6] In this sense a single page does constitute an adequate sample.

Second, although not usually presented as such, LALME localisation can be understood as the two-dimensional response variable in a regression model approximating the true relationships obtaining between a set of predictor variables grouped by text. The model is a model of a population of texts achieved by generalising from samples that are individual texts, and an individual text's localisation is its coordinates on the plane of fit that constitutes the model. The model was built from a sample of the population of texts written in less standardised orthography in England during the late medieval period (c. 1350–1450),[7] with the texts' orthographic forms as predictor variables. The model was initially trained on texts with known values for a single extra-orthographic

---

6 Around 2.3 of every 1,000 characters in the corpus (spaces excluded) is a token of the grapheme <r> in initial position, around 52.8 of them is a medial occurrence, and around 10.9 of them is a final one. An alternative to the present method of data collection is to examine either the same number of words from every text or a fixed set of words.

7 The exact date of production is unknown for the majority of the texts; the dates reported in LALME and the MEG-C "Catalogue of Sources" rely on secondary sources, including palaeographic studies, and may be approximate to within a quarter- or half-century.

variable added. The values for this variable, the coordinates representing a text's actual place of production in geographical space,[8] were considered values for the response variable for those texts. It was the model itself which determined the values for the response variable for the remaining texts. The model was developed dynamically, with the remaining texts being added successively to the training data and the model rebuilt after each addition.[9]

In other words, the variables considered were the orthographic ones (for every text) and the place of production (for some texts). Controlled for were each text's internal orthographic consistency, its date of production, and the amount of standardisation exhibited by its orthographic forms, in the case of all these three further variables on the basis of qualitative assessment performed prior to the building of the model. The procedure for determining internal consistency was described above. The 100-year span covered by the corpus was considered a homogeneous synchronic "time-slice", making date of production a constant,[10] and texts containing little dialectal "colouring" were altogether excluded. Also controlled for was a text's paleography in the sense that each text had to be written in a single hand for it to be included.[11] LALME terms a text fulfulling all these criteria as "a scribal text".

Criticism of LALME has centred either on the relationship between a scribal text's localisation and its actual place of production, since there are known cases of mismatches between the two;[12] or on the representativeness of a scribal text as a witness to the spontaneous usage of its scribe, since it is known that scribes partly introduced their own forms and partly reproduced forms from their exemplars when they copied texts (e.g., Millett 2012). There is reason to suspect that they did the same with allographs (Laing 1989; Varila 2014). However, both these two points of criticism seem misguided, since a scribe's spontaneous usage is but one of the variables selecting a scribal text's orthographic forms, and since the model's response variable purports to represent actual place of production only in the case of that subset of scribal texts for which values for it were known from the outset. What the model's response variable, localisation, is for all the remaining texts fulfilling the criteria described, is a numerical expression of the level of similarity between them in terms of values for all the predictor variables.

---

8 Or perhaps rather the place where the scribe was trained.

9 It is recommended in regression modelling to keep the training data and the test data strictly apart.

10 The corpus is skewed with the northern texts generally being younger than southern ones, and some texts' production date falls outside the 100-year span.

11 Cf. fn. 4 above.

12 A generic way of phrasing this criticism has become to point to the itinerant scribe with a stable dialect.

A scribal text's orthographic forms select further variables (and vice versa). Examples of such further variables are the text-type for the work it is a witness to and traditions surrounding that work, as well as the time it took to produce the specific copy, this copy's level of formality, its intended audience, and its physical dimensions, not to mention a variable mentioned in the previous paragraph: the extent to which its scribe reproduced orthographic forms from the exemplar or exemplars. These variables cannot be assumed to be perpendicular on each other, nor are they equally strong as predictors. They were not explicitly controlled for when the model was built. The localisation variable must absorb much of the variation these variables explain; but it is unknown how much of it the plane of fit leaves unexplained and to what extent the plane is collinear with any of them. However, for a scribal text that fulfills the other criteria described, it is only if a change in the value for any of these variables affects its internal orthographic consistency that the change will disqualify it as a single scribal text.

So, if paleography and orthography co-vary as theorised by McIntosh, the single-page sample will of necessity be representative of the scribal text with which it is associated since both will answer to identical values for the variables selecting forms. It is only if they do not co-vary that the sample will not be representative of the scribal text, in which case there will be a difference between them in the value for one or more of the predictor variables, which in turn will produce separate localisations for them. Plotting paleographic forms on the plane of fit representing scribal texts' LALME localisations is consequently a meaningful way of testing for the possible co-variation: if the paleographic forms are randomly distributed, there is no co-variation.

3. Corpus

To dwell on the possible predictor variables a little longer, several paleographers associate the tag on the round allograph of the grapheme <r> with formal grades within the Gothic Textualis and Cursiva family of scripts but they also note that its use does extend into informal grades (e.g., Derolez 2003: 84, 150). This is although the untagged form appears in formal as well as informal writings, even in de-luxe copies (Parkes 2008a: 124, figs 4.1 and 4.4), and it is although (or because) the tag itself is a non-cursive feature (cf. Derolez 2003). It is what Parkes (2008b) calls a subsidiary element—a non-essential element whose addition above or below cue-height increases production time but increases legibility and speeds up reading. I am not aware of any paleographic study demonstrating this association with formality. What evidence is laid out in support of it in the literature appears to be largely impressionistic—a series of examples extracted from selected primary sources

whose representativeness is asserted, addressed in vague, general terms, or not discussed at all.[13]

Other scholarship has called attention to two variables already mentioned, geography and text-type, as being particularly strong predictors of paleographic forms. LALME amply demonstrates that contiguous diffusion—the gradual spread of forms to geographically adjacent areas over time—explains the distribution of some S and W forms inasmuch as it shows the distribution of particular forms to be confined to contiguous areas on the plane of fit. This description fits the distribution of the use of the "y" allograph for the grapheme <þ>, which embodies a single contiguous area spanning the northern part of the plane.

By contrast, Doyle's (1994: 95, 96; emphasis original) "impression is that most styles and types are of long-established or quite rapid national dissemination [...] and of professional, not local, determination", and that they must have been "for long positively taught", including the use of "y" for <þ>.[14] Hector (1966: 12), Jenkinson (1915 [2014]: 6), and Parkes (2008b) follow suit, as does Benskin (1982: 16–17) by pointing out how "þ" is used for <þ> in legal and administrative documents associated with national government; such documents form a text-type excluded from his map. Diffusion of forms through professional networks, from one writing centre to another, will align with text-type. Crudely put, this is because smaller provincial centres saw scribes involved in copying several text-types for want of competitors, whereas there was sufficient demand in urban centres to allow scribes to specialise (Johnson & Jenkinson 1914: xvi-xvii). The mechanism is reminiscent of a proposed account of how uvular [r], supposedly originally a seventeenth-century Parisian pronunciation, came to have its present-day distribution in Western and Northern Europe (Trudgill 1974). That account holds that this phonological variant skipped from one urban centre to another, reaching more prominent and populous centres first and bypassing any rural area in between. Such hierarchical diffusion may result in what looks like a geographically contiguous

---

13 For example, Johnson & Jenkinson (1914: xiiv) seek to present to the student "a large, and on the whole fairly representative, series of examples of the kind of writing usually to be found in English documents from the Conquest to A.D. 1500" but they do not elaborate on their grounds for considering them representative.

14 Hector (1966: 12) writes that the paleographer "is gravely handicapped [...] by the fact that most of the hands written in English archives do not lend themselves to precise and confident classification by date and provenance. The self-conscious set hands which begin to appear in the 15th century are associated not with schools of handwriting in the cultural or local sense but with professional milieux and departments of administration; they are seldom seen except as the conventionally obligatory vehicles for formal documents of specific kinds. Outside such formal contexts the business hands of any one period may exhibit all the variety of which individualism is capable, and every document may be expected to illustrate in some degree the idiosyncrasies of its writer".

cluster of points on a distributional map but in fact rather represents a hierarchy of connections between the points with empty space in between them. It is especially hard—in fact, impossible—to determine the diffusion mechanism from maps giving isoglosses rather than individual points. It seems possible alternatively to understand the northernness associated with the use of the "y" allograph for the grapheme <þ> in this manner.

A further possibility is that the allograph's northernness results from both diffusion mechanisms (contiguous vs. hierarchical) operating in parallel. The distribution of vocalised /l/ in present-day Australian English shows them not to be mutually exclusive (Horvath & Horvath 1997). There is corroboration of this finding from tree-structured regression modelling, as this quantitative analytical methodology has demonstrated that geography (as represented by LALME localisation) and text-type select 16 other allographs than the one under study, with the former predictor variable being stronger than the other for some allographs and vice versa for other allographs (Thaisen 2017).

4. Results

I noted the presence or absence of tagged and untagged round <r> in 594 scribal texts, having disregarded 10 texts that were unlocalised in LALME. In doing so, I paid no attention to the proportion of one form to the other within a text, nor to small variations in execution between glyphs. I left out of consideration any text rendered in any other script than the one used for the body of the English text; other scripts are typically used for headings, phrases in Latin, or marginal annotations.

I took into consideration whether a given glyph occurred in initial, medial, or final position in the word, thus collecting a maximum of three glyphs of each of the two forms of the allograph per text. It was a compromise to distinguish only three levels: it was not practicable to record what mark respectively precedes and follows every recorded glyph, let alone record every glyph encountered in every text. I defined position in the word exclusively in relation to horizontal ordering of marks. This too was a compromise: occasional suspension of horizontal ordering characterises the Middle English writing system, inasmuch as both alphabetic marks and other linguistically significant marks may occasionally be placed above other marks.

It is the history of the round <r> allograph that motivates considering position in the word as a variable. Summing up secondary literature, the history is one of gradual relaxation of constraints on the allograph's occurrence. It is complementarily distributed with other allographs of the grapheme <r> in the basic ductus of Carolingian minuscule and at first also in that of this script's Gothic successors, being bound to the position after <o> and thus never

occurring word-initially. Over time, however, round <r> comes increasingly to compete with other allographs of <r> in positions following any alphabetic mark ending in a bow to the right—this distribution is one of the two "Rules of Meyer" defining Textualis. It is sometimes found also after <b> and <p> from the beginning of the 12th century (Parkes 2008a: 116), and after <h> and a particular allograph of <d> from the middle of the 13th century, all of which end in a bow to the right (Parkes 2008a: 124; cf. Johnson & Jenkinson 1914: 41).[15] Scribes tended, however, not to use the round <r> in final position in Textualis (Derolez 2003: 91). They did do so in Cursiva. The present corpus, which concentrates on the period from the mid-14th to the mid-15th centuries and in which Cursiva heavily outweighs Textualis, contains examples of round <r> in the contexts mentioned as well as in contexts following such a bow on <þ>, on the reversed and circular allograph of <e>, and on the allograph of <w> that is composed of two looped "l" and "3". Albeit examples are fewer, the present corpus also comprises ones of round <r> in positions which do not follow any bow to the right, including word-initial ones.[16] Other allographs of <r> may be found in most of these same contexts if not all of them.

Glyphs with the tag may derive from the mark "rum rotunda" employed in Latin to abbreviate the grapheme sequence <rum>, especially in the genitive case suffix "arum" or "orum" (Parkes 2008a: 119; Derolez 2003). The mark consists of round <r> with an elongated right limb crossed by a slanting stroke and makes its first appearance in the third quarter of the 12th century (Parkes 2008a: 119; Bischoff 1990: 135). The slanting stroke is the tag that indicates abbreviation. The present corpus contains examples of rum rotunda in Latin text (otherwise ignored). None of them appears to be identical to the tagged form of round <r> found in English text executed in the same scribal hand. The right limb tends to be less prolonged or not prolonged at all in such text, and the tag itself may attach to the stem or the right limb. It is not invariably executed as a separate stroke, as it may be a prolongation of the stem, with the right limb instead being executed as a separate stroke (Johnson & Jenkinson 1914: 42, exx. 16 and 19). The tag no longer has the abbreviating function but is a subsidiary element.

Five allographs of the grapheme <r> are recognised in the paleographical literature. They are "long r", whose descender dips below the baseline and may lack the return upstroke; "short r", which is right-shouldered and whose back does not dip below the baseline; "capital r", whose label is self-explanatory; "v-

---

15 Hector (1966: 57) does not find examples of round <r> after glyphs ending in a bow to the right but after "long-bodied letters".

16 Impressionistically, the word-initial examples first appear around 1400. McIntosh (1974: 54, n. 26) observes about the later Middle English period that tokens of round <r> (his "r3") do not exclusively occur after glyphs ending in a bow to the right.

shaped r", whose right-shoulder is attached at the base of the stem; and "round <r>" also known as "2-shaped r". The tag normally attaches to the round <r> allograph,17 but the corpus does house sporadic examples of what can best be described as short r with a tag. In Fig. 1 below, the three leftmost glyphs, randomly picked from the corpus, exemplify round <r> with a tag, while the two rightmost ones, taken from respectively Linguistic Profiles #0194 (left) and #5760 (right), would be examples of short r if the tag was absent, but examples of tagged round <r> if the right shoulder was absent.

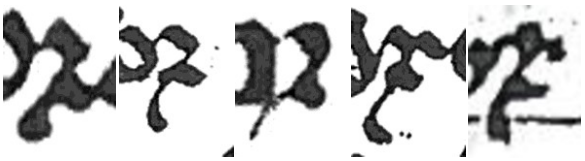

Figure 1: Examples of round <r> with a tag

I assigned values to four further variables in addition to position in the word and LALME localisation. They are all binary and combine to represent text-type in a fairly objective way, albeit crudely. What has preceded has argued the possible salience of text-type as a variable, which justifies annotating for it. The four variables record for each scribal text whether it is a document, whether it is written in prose, whether the poetic form is end-rhymed verse, and whether the verse form is alliterative. A document has the physical format of a document, which disqualifies a cartulary copy from counting as one. The two types of verse are self-explanatory, but non-versified text is classified as prose only if it is discursive; that is to say, tabularly arranged text such as a record of financial expenditure is considered neither verse nor prose.

5. Analysis

My methodology for determining the relative strengths of the variables as predictors of tagged round <r> was to grow a tree-structured regression model, specifically a conditional inference tree. This methodology, implemented in the "partykit" package for the R software environment for statistical computing, is designed to avoid overfitting a model to its training data and always grows the optimum tree for the specific variables (Hothorn & Zeileis 2018: 8). The methodology uses non-parametric significance testing to select predictors from

17 It is possible to define the tagged and untagged forms of the round <r> allograph as separate allographs. I follow standard conventions in paleography in not doing so.

among the variables and tends to exclude marginal predictors. A model whose accuracy exceeds 80 percent is likely to generalise to the population from which the training data were extracted, whereas a lower accuracy means the tree describes the training data only. The "partykit" package does not readily output a conditional inference tree's accuracy but it can be calculated for a corresponding logistic model.

The package outputs an upside-down tree comprising only the predictor variables required to explain the training data. The strongest variable appears nearest the root and the weakest ones form the terminal nodes. Every node represents a partitioning of the training data into exactly two sets at a value or level for a variable such that the sets are statistically significantly different from each other and maximally homogeneous. The weaker one among collinear variables becomes excluded in the process, and the requirement of homogeneity makes the partitioning process robust against outliers.

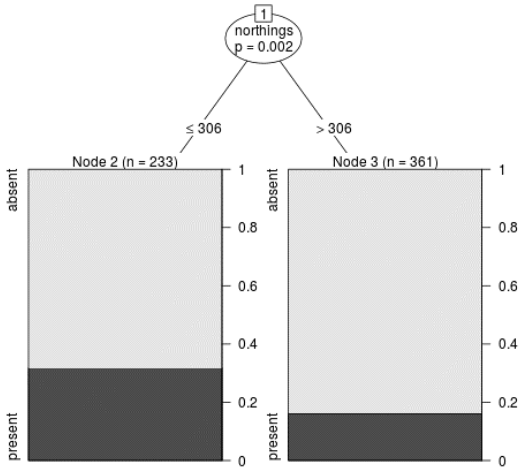

Figure 2: Conditional inference tree estimates for the relationship between predictors of the round <r> with a tag in medial position.

Note: The boxed number which appears above the name of a predictor at a node is an identifier, and the p value below it expresses the significance level for the partitioning of the training data at the given node. The values or levels at which the training data are partitioned are given along the edges, while an edge's length is immaterial. The n value which appears by a terminal node gives the number of observations in the training data conforming to the particular configuration of the predictors. A bar chart's left-hand scale gives levels of the response, in this case the presence or absence of the tag. A bar chart's right-hand scale gives the proportion of observations for each level according to the training data.

The conditional inference tree given in Fig. 2 estimates the relationship between predictors of the tagged round <r> in medial position, the variables tested being the two localisation ones plus the four text-type ones. It can be seen that LALME localisation in northings is the sole predictor. A scribal text localised below "the Midlands line" is statistically significantly more likely to contain the tagged allograph than one localised above it. The label describes a north-south division of the plane of fit falling at around 300 northings which is in evidence in the distribution of several other allographs too (Thaisen 2017). It also roughly marks the southern boundary of the use of the "y" allograph for the grapheme <þ> (Benskin 1982). The value corresponds to a line running across the Midlands from just north of Birmingham through Leicester and Norwich in geographical terms. No other tree is shown, since none of the variables predicts the tagged form in either initial or final position.[18]

However, adding the untagged form as a possible further variable revealed that it too is a predictor. There is a statistically significantly greater likelihood that a text will contain the tagged form if it does not also contain the untagged form. This is especially so in medial position for texts localised below the Midlands line (specifically, 343 northings). It would seem, then, that some scribes exclusively employed either the one form of round <r> or the other, whereas other scribes alternated between them, and that exclusive use of the tagged form is more likely in texts localisable to the south than the north.[19]

These findings might be argued to support the existence of co-variation between orthography and paleography. This is not only because the paleographical variation is interpretable when it is plotted against the texts' orthographic similarity to each other but also because many orthographic variables show a comparable north-south division of late mediaeval England; witness the dot maps for forms of THE, THESE, THOSE, THEY with initial <y>, THEM with medial <ai> or <ay>, the "hir" type for THEIR with simple <i> or <y> as the medial vowel, the "ech" type for EACH, the "eny" type for ANY (including "heny"), the "moch" and "mochel" types for MUCH, the "wol" type for WILL sg/pl with simple <o(o)>, the "til" for TO prep., etc.[20]

---

18 There are just three occurrences of the tagged form in initial position and twenty one of the untagged form.

19 None of the localisation or text-type variables predicts the untagged form in any of the three positions either, except if the tagged form is added to the equation as a possible predictor and then only in medial position. In that position, absence of the tagged form predicts the untagged form for texts localised east of 418 eastings. This value corresponds to geographical locations as far west as Salisbury and Swindon.

20 The electronic version of LALME is available online from https://www.lel.ed.ac.uk/ihd/elalme/elalme.html.

However, the counter-arguments weigh heavier. "Some correlation" between orthography and paleography is a more accurate wording than the stronger "co-variation", since the distribution of the round <r> allograph with the tag according to the localisation variable does not form the expected single contiguous area on the plane of fit but rather a scatter of points that vary in density. It may also point to a lack of co-variation that the text-type variables do not predict the tagged allograph, since a distinction such as prose vs verse is known to be salient with respect to orthographic variation. What has preceded has given theoretical reasons why it does not readily follow from the identification of localisation as a predictor that paleographical variants diffuse contiguously, and the distribution of the round <r> allograph with the tag in eastings and northings (not shown) is not such as readily to reveal the diffusion mechanism. What the study has identified by means of statistical testing is a focal interval of values for the northings variable where the tagged allograph occurs at a higher frequency, sometimes exclusively, and a transitional interval where its frequency is lower and where alternation with the untagged form is common. This distribution suggests the tagged form is spreading from the focal interval into the transitional interval, as the nomenclature implies.

6. Conclusion

All the while that the tree-structured regression models exemplified by Fig. 2 are the best possible account of the training data, they are not sufficiently accurate to ensure that they will successfully generalise to the population from which the training data were extracted. The logistic models corresponding to them, built by means of the "glm" function included in the "lme4" package for R,[21] have a C-statistic of .6111 for the model with LALME localisation in northings as the sole predictor of the tagged round <r> in medial position, one of .6655 for the model with northings and the untagged round <r> as predictors of it in medial position, and one of .5728 for the model with the untagged form as the sole predictor of it in final position. The C-statistic is a measure of the accuracy or goodness-of-fit for a regression model with a binary response variable. These values constitute a moderate fit.

It is reasonable to offer two reasons for the moderate fit. One is the inadequacy of the scales quantifying the variables. The levels "initial", "medial", and "final" overly crudely quantify the position-in-the-word predictor. It might be better represented by the shape of the preceding glyph, particularly whether it ends in a bow to the right, although there is no a priori

[21] The label "glm" is perhaps misleading, since it stands for "generalised linear model". The function builds a logistic model if the "family" variable is set to "binomial".

reason the tagged and untagged forms of the allograph should behave differently in this regard. Similarly, it is hard to test whether the text-type variable is a predictor in the absence of hard-and-fast, non-impressionistic criteria for what constitutes a text-type. The other reason is the partial applicability of the localisation variable in the first place owing to essential differences between orthography and paleography, as argued above.

Last, formality was not quantified so as to be able to test whether paleographers are correct to assert that this variable is a predictor. This is because, like the occurrence of untagged round <r>, formality is not a variable commonly invoked to explain orthographic variation in late Middle English, although register is salient in explaining variation at other levels of linguistic description, such as lexis. Several grading systems exist, as discussed by Derolez (2003: 13–27). It is the criteria for confidently separating the grades that constitute the particular point of disagreement. The minims are joined to each other in Cursiva scripts. While this requirement superficially seems a solid, objective criterion, the extent to which minims in fact are joined up often varies within actual scribal texts and may depend on the particular grapheme sequence; for example, whether the pen finishes on the baseline when the scribe executes a preceding grapheme's final stroke (cf. Derolez 2003: 123–130). Annotating for local context by collecting, say, one example of a medial round <r> with a tag in a "media" context and another in a "currens" context from the same scribal text would have impacted on the selection of modelling methodology.[22] Conditional inference trees and generalised linear models cannot handle the situation where, so to speak, some informants have answered a question more than once. The appropriate methodology would in that case have been one that allowed for "scribal text" as a random factor, such as the "lme4" package's "lmer" function.

But a more refined model focused on variables relevant to paleography is for a future study to develop. This paper's principal concern has been another: to address the possible co-variation between paleographical and orthographic variation suspected by McIntosh. To this end, it has employed quantitative analysis of quantitative data, where much paleographical literature is purely qualitative or relies on visual analysis of quantitative data, and rarely addresses whether the analysis possibly underfits or overfits the data. Although it is a fact that variation is present within and between scribal texts in the selection of allograph for the grapheme <r>, variables salient in explaining orthographic variation do not explain this allographic variation particularly well. There does not, therefore, appear to be strong support for co-variation; but the paper has found some measure of correlation, inasmuch as the distribution of a paleographical form is interpretable when it is related to the orthographic similarity between the texts in which it occurs.

---

22 This terminology for levels of formality follows Brown (1990: 1–2).

REFERENCES

Benskin, Michael. 1982. The letters <þ> and <y> in later Middle English, and some related matters. *Journal of the Society of Archivists* 7(1). 13–30. DOI: 10.1080/00379818209514199

Bischoff, Bernhard. 1990. *Latin palaeography: Antiquity and the Middle Ages.* Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511809927

Blake, Norman F. & Jacob Thaisen. 2004. Spelling's significance for textual studies. *Nordic Journal of English Studies* 3. 93–108.

Brown, Michelle P. 1990. *A guide to western historical scripts from Antiquity to 1600*. London: The British Library.

Derolez, Albert. 2003. *The palaeography of Gothic manuscript books: From the twelfth to the early sixteenth century*. Cambridge: Cambridge University Press.

Doyle, Anthony Ian. 1994. A palaeographer's view. In Margaret Laing & Keith Williamson (eds.), *Speaking in our tongues: Medieval dialectology and related disciplines*, 93–97. Cambridge: Brewer.

Hector, Leonard C. 1966. *The handwriting of English documents*. London: Edward Arnold.

Horvath, Barbara M. & Ronald J. Horvath. 1997. The geolinguistics of a sound change in progress: /l/ vocalization in Australia. *University of Pennsylvania Working Papers in Linguistics* 4(1). 109–124.

Hothorn, Torsten & Achim Zeileis. 2018. *A toolkit for recursive partytioning*, version 1.2–1. https://cran.r-project.org/web/packages/partykit/partykit.pdf (accessed 25 May 2018).

Jenkinson, Hilary. 1915. *Palaeography and the practical study of court hand.* Cambridge: Cambridge University Press.

Johnson, Charles & Hilary Jenkinson. 1914. *English court hand, A.D. 1066 to 1500*. Oxford: Clarendon Press.

Laing, Margaret. 1989. Linguistic profiles and textual criticism: The translations by Richard Misyn of Rolle's *Incendium Amoris* and *Emendatio Vitae*. In Margaret Laing (ed.), *Middle English dialectology: Essays on some principles and problems*, 188–223. Aberdeen: Aberdeen University Press.

McIntosh, Angus. 1974. Towards an inventory of Middle English scribes. *Neuphilologische Mitteilungen* 75(4). 602–624.

McIntosh, Angus. 1975. Scribal profiles from Middle English texts. *Neuphilologische Mitteilungen* 76(2). 218–235.

McIntosh, Angus, Michael L. Samuels & Michael Benskin (eds.). 1986. A *linguistic atlas of Late Mediaeval English*. Aberdeen: Aberdeen University Press.

Millett, Bella. 2012. Scribal geography. In Elaine Treharne, Orietta Da Rold & Mary Swan (eds.), *New medieval literatures 13*, 183–197. Turnhout: Brepols.

Parkes, Malcolm B. 2008a. Handwriting in English books. In Nigel J. Morgan & Rodney M. Thomson (eds.), *The Cambridge history of the book in Britain.* Vol. 2: 1100–1400, 110–135. Cambridge: Cambridge University Press.

Parkes, Malcolm B. 2008b. *Their hands before our eyes: A closer look at scribes*. Aldershot: Ashgate.

Robinson, Peter M. W. & Elizabeth Solopova. 1993. Guidelines for transcription of the manuscripts of the Wife of Bath's Prologue. In Norman F**.** Blake & Peter M. W. Robinson (eds.), *The* Canterbury Tales Project *occasional papers volume I*, 19–52. Oxford: Office for Humanities Communication

Stenroos, Merja, Martti Mäkinen, Simon Horobin & Jeremy Smith (compilers). 2011. *The Middle English grammar corpus*, version 2011.1. Available at https://www.uis.no/mest [accessed 3rd July 2016].

Thaisen, Jacob. 2011. Adam Pinkhurst's short and long forms. In Jacob Thaisen & Hanna Rutkowska (eds.), *Scribes, printers, and the accidentals of their texts*, 73–90. Frankfurt: Peter Lang.

Thaisen, Jacob. 2017. Secretary letter-shapes in County Durham. *Folia Linguistica Historica* 38. 263–280. DOI: 10.1515/flih-2017-0009

Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 3(2). 215–246. DOI: 10.1017/S0047404500004358

Varila, Mari-Liisa. 2014. Graphetic variation within one scribal hand as evidence on manuscript production. *Studia Neophilologica* 86s1. 157–170. DOI: 10.1080/00393274.2013.834107