Thaisen, J. (2014) Initial position in the Middle English verse line.
***English studies,* 95(4), pp. 500–513**

Link to **published article:**
DOI:10.1080/0013838X.2014.924275
(Access to content may be restricted)

Initial Position in the Middle English Verse Line

Jacob Thaisen

This paper establishes that spelling forms collected from initial position in the Middle English verse line have unique characteristics, and it discusses why this is so. The paper first addresses scribal copying practices, before describing the utility of letter-based N*-gram models in objectively comparing scribal copies in terms of their spelling. Testing of models trained on a corpus totalling ten manuscripts demonstrates that initial position regularly prompted scribes to suppress their tendency to introduce their own spelling forms in favour of replicating those encountered in their exemplars. The discussion attributes this behaviour to the operation of two mechanisms. One mechanism is psycholinguistic in origin, while the other is rooted in manuscripts' production and so implies a codicological dimension to spelling variation.*

Something is different about initial position in the verse line. Scholars seasoned in compiling spelling profiles from Middle English manuscripts will nod in acknowledgement. It will be their experience, like it is my own, that spelling forms collected from this position often stand out in a profile. The observation merits empirical verification and a testable explanation.

It was a deliberate strategy on the part of late medieval English scribes not invariably to aim at producing a carbon copy of their exemplars in respect of spelling forms. Scribes felt free to introduce their own spelling forms into the copy they were producing. The reason that departures from the exemplars were possible was the absence of normative spelling conventions during the period. The writing system was variable with no individual spelling form regarded as the standard or canonical form, and it was by no means unusual for a scribe to command more than one spelling form for any one word. Nevertheless, it was not an unchecked process that saw scribes effect these departures.

Two types of mechanism are recognised in the literature as having led to a suspension of the "translation" process. The first mechanism responds to properties of the verse being copied. It operated in cases where replication of a spelling form from the exemplar was essential for maintaining rhyme, metre, or alliteration, or, at a more general level, for complying with possible conventions of the text or its genre. The checking effect brought about by a desire never to upset these properties was presumably strongest in those cases where the scribe's own spelling form suggested a phonological realisation different from that suggested by the form found in the exemplar at the given location in the text, although the veracity of this presumption deserves empirical testing; it has, for example, been argued that eye-rhyme is intended in some Old and Middle English verse.[1] The effect was so strong as to separate the spelling practices adopted for line-final position from those characteristic of other positions in the verse line—indeed, a comprehensive account of the interaction between a scribe and his exemplar once deemed it "not unreasonable to speak of scribal diglossia".[2]

What may be called priming was the second mechanism at work. A psycholinguist would in relation to the copying situation understand this notion as the act of biasing a scribe subconsciously to select spelling forms he has encountered previously. The individual form could rank among the scribe's (possibly several) forms in active use for the word in question, or it could fall outside this range, but none the less be familiar to him from his exposure to the written word, including by way of the exemplar he was currently copying. Only rarely was the reproduced form altogether foreign to the scribe.[3] The end result may be promotion ("entrenchment") of a form to make it the scribe's default selection.[4] The checking effect brought about by priming consequently led to a spelling profile skewed in the direction of the exemplar and only partially representative of the scribe's unchecked usage.

Widely acknowledged in informal conversation, but none the less under-researched, is a possible third mechanism which also may have constrained a scribe when selecting among the spelling forms available to him. The mechanism is associated with initial position in the verse line, but may conceivably have operated anywhere in a text, especially in connection with notable words or phrasal boundaries. It is the replication of spelling forms from the exemplar in order for the scribe to rely on them as a means

[1] Stanley.

[2] Benskin and Laing, 70.

[3] It has been shown to be hard to implement deliberate "translation" into any other usage than one's own. The evidence comes from authorship attribution studies of modern materials. For example, Victorian novelist Charles Kingsley's novel *The Tutor's Story* was completed posthumously by Lucas Malet, his daughter Mary's pseudonym. Malet states in the preface that she has tried to preserve peculiarities of her father's style; yet David L. Hoover narrates how tests based on lexical choices confidently separate them and a gradual transition is in evidence. It similarly appears to be possible to measure the extent of Jeremiah Curtin's wife's contribution to what was published as his translations, as shown by Jan Rybicki.

[4] It may be possible to explain in terms of incipient entrenchment the known examples of a scribe's bias toward the exemplar strengthening in step with his exposure to it. The present materials contain such examples. Jeremy J. Smith suggests that priming may account for the presence of the spelling form <oughne> OWN (adjective) in manuscripts Corpus and Harley 7334, as their shared scribe may have entrenched this form through repeat copying of Gower texts hosting the related form <ougne>. For a case of weaker subconscious adoption of spelling forms from an exemplar, see Blake and Thaisen. These authors show that the scribe of the Christ Church manuscript accepts what must have been old-fashioned spelling forms to him, such as <nat> NOT and the inflectional suffix <e>, increasingly as his copying proceeds.

for him to avoid eye-skip. Eye-skip is the phenomenon of a scribe losing his place in a text, as he constantly turns back and forth between the exemplar he is consulting and the copy he is producing. The scribe may, as a result, copy a passage twice or leave another out, necessitating correction, which may possibly upset the presentation.

Frances McSparran appears to be the first scholar explicitly to have proposed reliance on spelling forms as possible finding tools. She did so in a discussion of spelling forms present in the English sections of British Library, Harley 2253. However, while it is unambiguous that the forms included in McSparran's profile do cluster according to postion in the verse line, she offered no evidence in support of the proposal other than a foot-note stating how collecting data for her study of the miscellany had "reinforced my [McSparran's] belief in the important function of the first word in a line of verse as a finding tool."[5] Nor did she consider alternative explanations for their clustering.[6]

In what follows, I provide empirical support for the retention of spelling forms from the exemplar in line-initial position. I do so by measuring how similar spelling forms collected from initial position are to those found elsewhere in the verse line. My metric, described next, is perplexity of letter-based N-gram models. The corpus comprises diverse verse texts in diverse scribal hands. Their diversity means that they "control" each other in various ways to permit generalisation about the spelling of scribal copies from any pattern visible in the similarity metrics, which are presented following a description of the experimental set-up. A pattern is evident, and a discussion of possible explanations why that pattern should be the norm in Middle English manuscripts lends credibility to McSparran's proposal in combination with priming. I suggest future means of verifying the proposal by way of negative evidence before concluding in the final section.

N-gram models make it possible to measure how similar Middle English texts are in terms of spelling, despite their lexical differences.[7] An N-gram is straightforwardly a letter sequence of length N—<a> is a 1-gram, <ab> a 2-gram, <abc> a 3-gram, and so forth—and a model is an exhaustive inventory of the N-grams that occur in a training text together with their frequencies. Every N-gram found in a test text receives a separate probability estimated from its frequency according to the model plus a weighting. The log-averaged inverse of these probabilities is a model's "perplexity": an objective quantification, always in the shape of a positive number larger than one, of how well it predicts the test text; a low value indicating a great similarity.[8]

The N-gram modeller views lexical difference between texts as a sampling effect and reduces it by applying routine techniques from natural language processing designed for the purpose, specifically smoothing and interpolation. The term smoothing

[5] McSparran, 399, fn. 25.

[6] Eugène Vinaver does not address the possibility in his discussion of the workings of eye-skip and other mechanical errors and their consequences for the scribal copy of a text.

[7] See Fink and references cited there for an introduction to N-gram models.

[8] The nature of the computations is such that in the comparison of a model and test data, there is no immediate way of extracting the specific grams which discriminate the best.

describes techniques for assigning less relative weight to grams frequent in training data and more relative weight to infrequent ones, including unattested ones. This is achieved by adding to the recorded frequencies. Since the weights are determined by patterns observed in the training data, unattested grams do not receive uniform probability. A linearly interpolated model for the gram length N is a model which also contains one for the gram length N-1, which in turn contains one for the gram length N-2, and so forth. Any N-gram unattested according to a model is recursively dissolved into its two constituent (N-1)-grams, which may be attested. The volume of training data consequently has a bearing on the accuracy of the individual probability estimates but only trivially increases or decreases their average and so also a model's perplexity. The greater perplexity associated with line-initial position cannot for these reasons be attributed to short words like AND, OF, THE and WHEN occurring with above-average frequency in it, including in their abbreviated spelling forms. Nor is it attributable to avoidance of the letter <þ> in initial position, or to a greater number of words and grams having been collected from medial position than from the other four positions.[9]

The Experiment

To be able to construct models and compute their perplexity for verse-line positions against one another, I obtained electronic transcripts of two sets of monolingual Middle English materials. What dictated my selecting the sets was that their diversity would allow me to control for any possible scribal, authorial or textual idiosyncrasy. The first set, transcribed at the University of Sheffield, consisted of nine practically complete manuscript copies of Geoffrey Chaucer's poem the *Canterbury Tales*, from which I took out the prose tales of Melibee and the Parson so that only versified text remained. The manuscripts were Aberystwyth, National Library of Wales, Peniarth 392 D ("Hengwrt"); Cambridge, University Library, Dd.4.24 and Gg.4.27; London, British Library, Additional 35,286, Harley 7334, and Lansdowne 851; Oxford, Christ Church, 152; Oxford, Corpus Christi College, 198; and San Marino, California, Huntington Library, El.26.C.9 ("Ellesmere"). Each of these manuscripts was copied

[9] This can be verified experimentally. Figures 1 and 2 illustrate a pattern of similarity between verse-line positions. This pattern, a U-shaped curve, is also in evidence with training and test data containing an equal number of 3-grams exclusively representing the initial letters of spelling forms. Moreover, there is no direct relationship between models' perplexity and spelling forms' length. To see this, consider that the average shortest spelling forms occur line-initially, but the average longest ones in the other position associated with great perplexity, line-final position. Finally, the interest in <þ> arises from scribes of late medieval English manuscripts regularly disfavouring this letter in line-initial position, perhaps for aesthetic reasons. However, the distribution of a single 1-gram should in theory be of negligible consequence in an N-gram model, especially in a smoothed model. Examples of line-initial <þ> number as few as one in the Additional 35,286 manuscript, two in the Christ Church manuscript, none in the Ellesmere manuscript and one in the Hengwrt manuscript, while more than 11,000—roughly one in five—lines in the Auchinleck materials start with this letter. The U-shaped curve differs little between the manuscripts despite this uneven distribution of <þ>, thus confirming this inconsequentiality in practice.

in its entirety by a different scribe, although two of the scribes each copied two of the manucripts. They range in date from around the turn of the fifteenth century to that century's third quarter. Their textual interrelationships are complex but none of the manuscripts is considered to be a direct copy of another, and it is debatable to what extent any pair of them share immediate exemplars in whole or in part. All nine manuscripts may descend independently from Chaucer's draft materials for the poem. The *Tales*' nearly 20,000 lines of verse vary in their poetic form, although the bulk of the text is universally accepted as being by a single author.

The second set of materials, obtained from the Oxford Text Archive,[10] is contained in a single volume but spans a larger set of variables. It is Edinburgh, National Library of Scotland, Advocates' 19.2.1 ("Auchinleck") with forty-four texts executed in six scribal hands, a change of hand always falling at a textual boundary and scribes 1 and 3 each contributing more than one stint; of these contents, I ignored scribe 2's stint: an exclusive listing of Norman baronial surnames laid out with one surname per line. The remaining forty-three texts are versified and vary in their form. Scribe 1 is responsible for upwards of two-thirds of the c. 58,000 extant lines and may have coordinated the compilation of this now-defective volume, which dates from the second quarter of the fourteenth century. Several of the texts survive uniquely or in their earliest known copy in the Auchinleck manuscript but more than one author or translator are evident.[11]

My toolkit took the transcripts of the *Canterbury Tales* as training data and built a separate, linearly interpolated model for each of the gram lengths 1–5 for each of five verse-line positions for each manuscript. Smoothing of the models proceeded according to the method devised by Ian Witten and Timothy Bell, which lets the weight of a gram be determined by the number of unique contexts in which it is attested.[12] The verse-line positions distinguished were initial, second, medial, final-but-one and final, with space employed as the separator between spelling forms occupying these positions and case distinctions levelled.[13]

[10] University of Oxford Text Archive; the materials are the deposited source files for Burnley and Wiggins, eds.

[11] Models trained on the one set of materials cannot be tested on transcripts from the other set or vice versa for lack of direct comparability between them. Their incompatibility is a result of the sets' having been transcribed and edited according to different protocols. For the transcription protocol for the *Canterbury Tales* manuscripts, consult Robinson and Solopova; for the principles followed in preparing the Auchinleck manuscript, consult Burnley and Wiggins, s.v. "Editorial and Transcription Policy." Differences in the editorial treatment of spaces is the probable reason that the results reported below for the verse-line positions "second" and "final-but-one" show them to align more closely with "medial" position in the Auchinleck materials than in the *Canterbury Tales* materials.

[12] Witten and Bell.

[13] The number of forms or syllables to a line is not constant in the present materials. In partitioning the transcripts, I none the less considered a form as medial if it occupied none of the other four positions. Concretely, I extracted from the transcripts as training data for medial position every line less its initial, second, final-but-one and final forms, and subsequently converted these data for medial position into a "bag of words" to eliminate any possible effects of including grams spanning word boundaries. I accepted, as a single exception, four-word lines as lines containing no medial word, but ignored all lines containing still fewer words (such lines are usually defective because of physical damage to the folio).

The toolkit, the *SRI Language Modelling Toolkit*,[14] next took the same transcripts as test data. It returned, separately for each manuscript and for each model trained on that same manuscript, a separate perplexity on the test data for each verse-line position. It never computed the perplexity of any model trained on data from one manuscript on test data collected from another manuscript.

I repeated this procedure with the Auchinleck transcript as test data but only for the gram length 3 and with the transcript divided into seven equal-sized segments for each verse-line position. All segments consisted of consecutive lines (as opposed to alternate or random lines), and segmentation paid no attention to scribal, textual or codicological boundaries. I maintained these segments when the transcript subsequently served as test data.

Results

Figure 1 comprises, respectively for each gram length, a boxplot of the mean perplexities obtained for each verse-line position with the nine *Canterbury Tales* manuscripts. In all five plots, which are due to the *R* software environment for statistical computing, the vertical axis gives perplexity, while the horizontal axis gives position in the verse line. A box ends at the twenty-fifth and seventy-fifth quartiles with the horizontal line inside a box marking the statistical median. A T-bar at each end extends one and a half times the interquartile range from the median, and circles represent outliers.

As is apparent from Figure 1, the boxes always describe a curve with a shape reminiscent of a capital letter U with a comparatively short left arm. A one-way ANOVA and Tukey's HSD, conducted by means of *R*, show whether the mean perplexities obtained from the manuscripts for the five verse-line positions significantly differ from one another (significance defined as $P < .050$). They do for all positions with 4- and 5-grams (all $P < .030$). With 3-grams, they again do (all $P < .001$), except for the pair "second and final-but-one" ($P = .267$). With 2-grams, the pairs "second and medial" ($P = .074$) and "second and final-but-one" ($P = .439$) differ non-significantly in terms of their mean perplexities, while all other pairs, including "medial and final-but-one", differ significantly (all $P < .001$). With 1-grams, only a few position pairs show significant differences: "medial and final" ($P < .030$), "initial and medial" ($P < .040$), "second and final" ($P < .040$), "final-but-one and final" ($P < .050$) and "initial and second" ($P < .050$). The remaining pairs show non-significant differences.

In other words, the number of populations decreases with gram length: 4- and 5-grams give five populations, 3-grams four, 2-grams three and 1-grams one—2- and 3-grams best permit generalisation about spelling, since longer grams more closely reflect the lexicon. There are patterns in how the mean perplexities for the five positions group. They suggest that the positions "second" and "final-but-one"

[14] Stolcke. The toolkit is downloadable from http://www-speech.sri.com/projects/srilm/ (cited 13 June 2012).

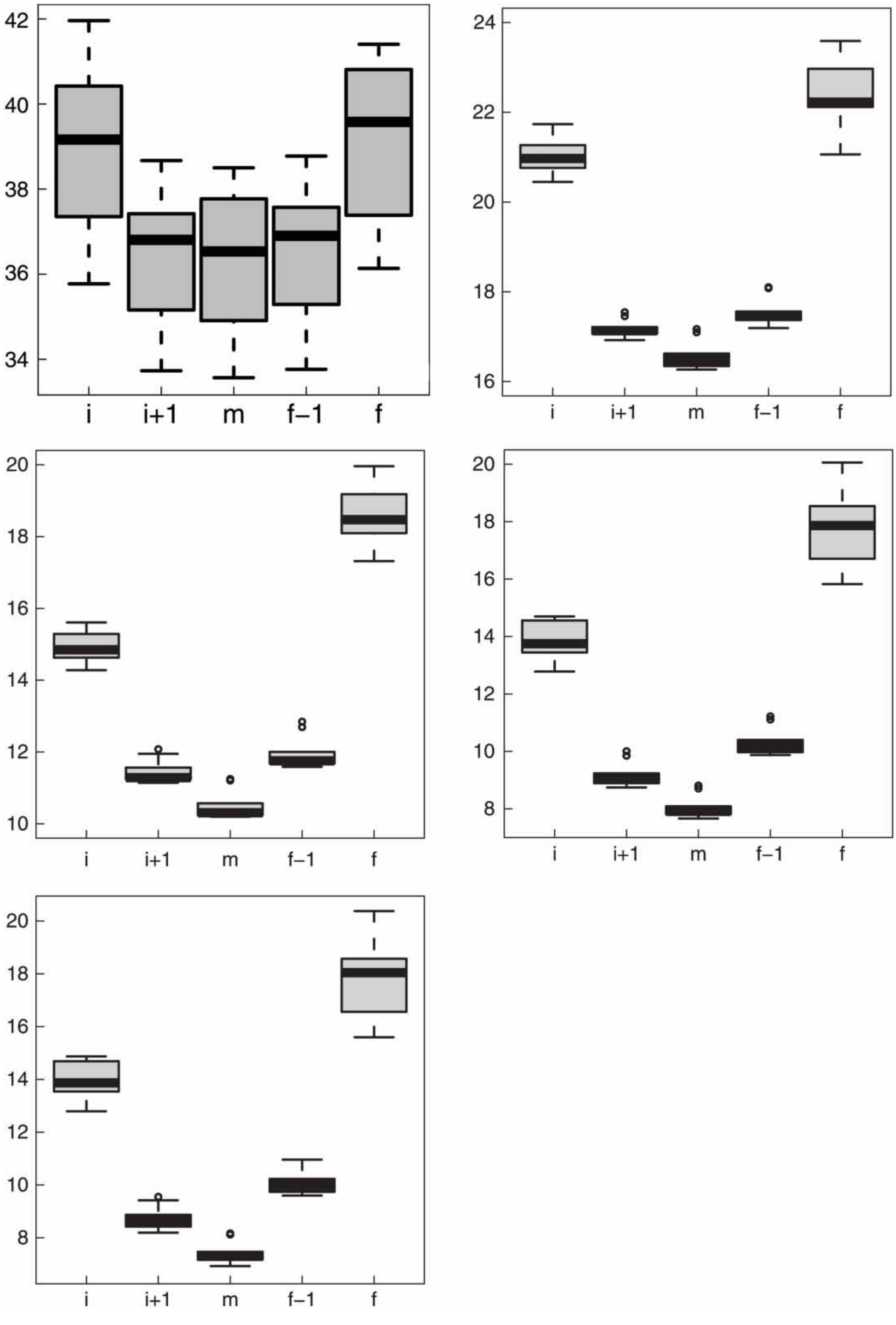

Figure 1 Mean perplexity by position in the verse line for models of nine manuscripts of the *Canterbury Tales*. The gram lengths are 1 (top row, left), 2 (top row, right), 3 (middle row, left), 4 (middle row, right) and 5 (bottom row, left). The positions distinguished are initial ("i"), second ("i+1"), medial ("m"), final-but-one ("f-1") and final ("f").

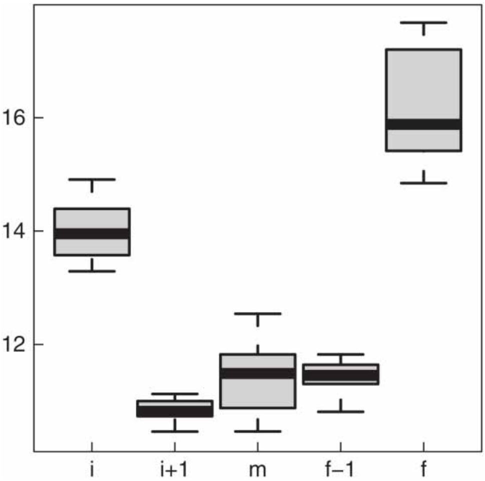

Figure 2 Mean perplexity by position in the verse line for 3-gram models of the auchinleck manuscript. The positions distinguished are initial ("i"), second ("i+1"), medial ("m"), final-but-one ("f-1") and final ("f").

represent similar populations, which in turn are similar to the population sampled in medial position. The positions "medial" and "final" mark extremes which initial position falls between.

Figure 2 offers a boxplot of the mean perplexities computed for the Auchinleck materials. It has the layout familiar from Figure 1, and it can be seen that the boxes describe a U-shaped curve similar to that evidenced by that previous figure. Tukey's HSD on the results of a one-way ANOVA distinguishes three populations, since the pairs "second and medial" (P = .357), "second and final-but-one" (P = .347) and "medial and final-but-one" (P = 1.000) do not differ significantly from each other.

Discussion

The U shape of the curve indicates that something leads to recurrent suspension of the scribal "translation" policy in line-initial position. What justifies this conclusion is that final position is uncontroversially recognised as the location in which scribes most frequently suspended this policy in favour of replicating spelling forms from their exemplars. The resulting blend of forms is consistent with the comparatively great perplexity also in evidence in initial position. At the same time, the diversity of the *Canterbury Tales* and Auchinleck materials dismisses scribal, authorial or textual idiosyncrasy as the reason for the curve's shape and shows that it is a widespread characteristic of scribal copies of medieval English verse based on written exemplars.

The first of the two previously recognised constraining mechanisms may likewise be dismissed. This is because among the properties of verse known possibly to check a scribe in his selection of spelling form, it is only rhyme which is associated especially strongly with any specific position in the verse line. Rhyme is, of course, not a feature of beginnings of verse lines.

More promising is the second mechanism, priming. It is known from psycholinguistics that what primes is a web of factors relating to a form's graphotactic and semantic complexity in addition to its context, frequency of occurrence and, recursively, the time elapsed since its latest occurrence ("lag").[15] These factors are under-researched in relation to scribal copying practices, historical linguists having concentrated on devising other kinds of methodologies for how to identify a form as exemplar-derived, such as by dividing a set of spelling forms into geographically incompatible subsets. Even without a firm evidential basis, it none the less seems safe to infer from studies of other materials that when a form primed a scribe, it did so because it was somehow distinctive.

Litterae notabiliores satisfy the condition of distinctiveness. These are distinct or regular letter-shapes made notable through enlargement, and often touched with or executed in a different coloured ink. A modest one is archetypically found at the beginning of a smaller textual unit such as a line or stanza, whereas one opening a larger unit such as a chapter or book is correspondingly larger and may sometimes constitute a miniature painting. Such hierarchies characterise the present manuscripts too, although the degree of elaborateness varies. The Christ Church manuscript has the least elaborate lowest-level litterae notabiliores, with many letter-shapes at best notable exclusively through enlargement, while the most elaborate ones are those found in the Auchinleck manuscript. The latter's folios 16^{vb} and 176ra are typical pages in this regard, respectively executed by its scribes 1 and 5. Unlike in any of the Chaucerian materials, the individual littera notabilior is on both pages not only picked out in coloured ink but also set off by a wide space from the remainder of the line to which it belongs; the latter device is supported by ruling.[16] Higher-level ones are numerically too few in any of the manuscripts to impact on the perplexity metrics. Despite the variation between the manuscripts in how letter-shapes signal their structural hierarchies, the special effort required to produce them must universally have focussed attention on the left edge of lines.[17]

It did not, however, focus the scribes' attention exclusively on the litterae notabiliores themselves. This can be stated with confidence, since the U-shaped curve remains substantially unchanged when the first letter of every word is removed from the training and test data. What is replicated from the exemplar, then, may be as much as the first word in its entirety, rather than its first letter.

[15] For an overview, see Traxler and Gernsbacher, eds., and references cited there, esp. chapter 10.

[16] Pearsall and Cunningham, eds., xiv schematically illustrate the various ruling patterns in the Auchinleck manuscript.

[17] The stemmatic position of the present manuscripts is such that no direct evidence is available to determine whether the exemplars contained any notable letter-shapes. It seems reasonable to conjecture that regular lines in them opened with a littera notabilior of some kind, although it is possible that some exemplars were working drafts—this is of course especially the case with the *Canterbury Tales*, since Chaucer left his poem unfinished. It is conceivable that not all lines were written out in full in such drafts.

An additional something that may seem to deserve further investigation is possible priming by the left edge's heading a segment of text to be copied. On the one hand, it is known from psycholinguistics that conscious memorisation and retrieval of a series of letters or words follow the direction of reading, unless the reader encounters processing difficulty. It is also known that accuracy in retrieval decreases with the length of the series, although not at a constant rate.

It may, on the other hand, be little more than presupposition to maintain that a scribe should also subconsciously focus on the beginning of a segment to be copied. The reasons that such a focus cannot be presupposed are not only that it strictly is unascertained where segments began and ended and that exemplars were always available for consultation. It is also that the eyes of a proficient reader do not move linearly across a line of text: they alternate between short, rapid movements ("saccades") and resting points ("fixations"). Extraction of meaningful information happens during fixations, but a reader's perceptual span is skewed in the direction of reading. A reader of English, which is read from left to right, picks up fewer letters to the left of a point of fixation than to the right and frequently skips short words altogether.[18] Most concretely, however, Figures 1 and 2 show changes in mean perplexity in the verse line too abrupt to be attributable to any realistic decrease in accuracy of serial retrieval, be it conscious or subconscious. This abruptness agrees better with a strictly local constraint operating on line-initial position.

A third mechanism must, therefore, combine with priming to make a scribe regularly copy literatim, from the beginning of a line in an exemplar, a series of letters up to one word in length. Eye-skip was an unfortunate lapse in execution, since it spoiled the integrity of a text or, if remedied, possibly also the neatness of its presentation. The potential for the lapse to occur could be reduced through a scribe's relying on finding tools to help him navigate between the exemplar he was consulting and the copy he was producing. I suggest, following McSparran, that the exemplar-derived spelling forms found line-initially are such unobstrusive tools. If this was their function, it was consciously that a scribe furnished his copy with them, just like the replication of rhyming words must have been.[19]

Is it possible to verify the habitual scribal employment of line-initial spelling forms as finding tools? It would bring no clarification to study correspondences between a copy and its exemplar or between two copies made from the same exemplar. Such a study would confirm the U-shaped curve but would not readily reveal why it has this shape. Relatedly, metrics such as perplexities obtained with N-gram models only indicate how similar texts are, but can never in themselves explain why they are similar. The present methodology may none the less bring indirect verification, if in fact no U-shaped curve results from applying it on versified materials where it was indisputably extraneous to rely on line-initial spelling forms as finding tools.

[18] See Staub and Rayner for a description of eye movements during reading.

[19] It seems logical to suggest that the function as a finding tool would be best fulfilled not by replicated spelling forms but by replicated letter-shapes; the present study was not designed to investigate this possibility.

Neither materials copied to dictation nor the authorial holograph satisfies this condition, while the amateurish product may. Consider the first of these classes. Its members may be fair copies produced from notes taken to dictation rather than the notes themselves, which would have constituted a written exemplar of sorts and would have recorded the verse lines on ephemeral carriers such as wax tablets or parchment scraps. It is conceivable of such notes that they would have contained a distinct set of spelling forms compared to the fair copy, since the nature of the carrier itself would have prompted the scribe to select the shortest possible forms from among those available to him. Whether notes did intervene or not, the greater obstacle to verifying the use of spelling forms as finding tools by means of materials belonging to this class is, however, that it appears to have very few certain members as far as longer versified texts in English are concerned. The visual impairment afflicting John Audelay perhaps makes him the most obvious candidate for a dictating poet, but scholars have none the less suggested that his two scribes relied on written exemplars for at least some of the texts in his manuscript, Oxford, Bodleian Library, Douce 302.[20]

A similar argument applies to the second class: the authorial holograph. It is a moot point to what extent the production of an authorial holograph entailed visual consultation of an exemplar and if it did, whether the process of the holograph's production should be regarded as analogous with the typical process of a manuscript's production involving no coincidence of author and scribe. Although the term authorial holograph does suggest stemmatic primacy, the image hardly convinces of an author composing a longer versified text directly in fair copy, and an author's spelling forms are not necessarily constant across media. As with the first class, it is conceivable that an authorial draft sketched out on wax tablets or other ephemeral media contained a distinct set of spelling forms.

The case in which the modal verb "may" applies relates to copying proper. It is that of the copy prepared by someone able to write yet unfamiliar with or disregardful of the technique of relying on spelling forms as finding tools. Such a description could fit an otherwise practised writer unaccustomed to copying longer texts, let alone entire codices. Promising for this reason are commonplace books like those respectively associated with the London grocer Richard Hill (Oxford, Balliol College, 354) or the provincial aristocrat Humphrey Newton (Oxford, Bodleian Library, Lat. misc. c. 66, "Capesthorne"), although the comparatively small amount of copied verse contained in them may constitute an insufficient basis for conclusive verification. Andrew Taylor's search for "manuscrits de jongleur" cast serious doubt on their existence as a separate class, but it did result in a shortlist of both rolls and manuscripts if

[20] See Audelay. Six of the carols and the verse sermon "Virtues of the Mass" appear in similar or variant forms in other manuscripts, two texts are translations from Latin, and another text is an excerpt from Richard Rolle's *Form of Living*. In addition, John Audelay's frequent use of anaphora may have rendered initial position useless for the placing of a finding tool; for example, numerous consecutive lines begin with *haile* AVE or *O Ihesu* OH JESUS (variously spelled) in "Salutation to Christ's Body," " Salutation to Jesus for Mary's Love," as well as certain devotions to Mary.

not written out to dictation, then at least less carefully planned, quickly executed and never intended to be quality products.[21] To the extent that these items contain copied or dictated verse, they carry the best promise of verification.

Conclusion

To sum up, both peripheral positions in the verse line constrain a scribe in his selection of spelling form. The constraints are non-categorical and more local in nature than has perhaps been realised. Both positions record an above-average number of spelling forms per word, the sources for the forms being sometimes the scribe and at other times the exemplar. A form found in a non-peripheral position in a scribal copy conversely has an above-average likelihood of being representative of the scribe's unchecked usage. Moreover, under the hypothesis that the spelling forms found in a peripheral position served as inconspicuous finding tools, the distribution of spelling forms within a scribal copy affords glimpses into how the copy was produced. These findings have the implication for students of English medieval texts and their manu-scripts that spelling deserves a more prominent place in both codicological and textual studies. A specific implication of interest to the historical linguist relates to the distance between a scribe's own spelling forms and those present in his exemplar. For a peripheral position, this distance does not appear to be the principal factor dictating to a scribe whether to insert his own spelling form or replicate the one found in the exemplar. That factor is a codicological one.

Lastly, the present study has demonstrated how adequate the perplexity of N-gram models is as an objective similarity metric for Middle English spelling data. Such models are a fixture in natural language processing. They have, however, rarely been constructed for the variable spelling systems characteristic of Middle English, most likely because a successful model presupposes a sizable body of training data. The tradition has instead been for the researcher to assess similarity based on visual, predominantly qualitative comparison of spelling forms of selected words collected from samples of texts. Diplomatic transcripts of longer medieval English texts are increasingly becoming available in electronic form to serve as training data. Their arrival promises full models optimised through smoothing and interpolation as a basis for rigid testing.

Acknowledgements

The author thanks Joanna Nykiel and Merja Stenroos for their constructive remarks on earlier drafts. Versions of the paper were presented to the International Congress on Medieval Studies, Kalamazoo and the International Conference on English Historical Linguistics, Zürich in 2012. The author was the grateful beneficiary of a stipend from

[21] Taylor.

the Medieval Institute at the University of Notre Dame, which facilitated the preparation of the paper.

References

Audelay, John. *Poems and Carols (Oxford, Bodleian Library MS Douce 302)*. Edited by Susanna Fein. Kalamazoo, MI: Medieval Institute Publications, 2009.

Benskin, Michael, and Margaret Laing. "Translations and Mischsprachen in Middle English Manuscripts." In *So Meny People Longages and Tonges: Philological Essays in Scots and Mediaeval English Presented to Angus McIntosh*, edited by Michael Benskin and Michael L. Samuels, 55–106. Edinburgh: The Middle English Dialect Project, 1981.

Blake, Norman F., and Jacob Thaisen. "Spelling's Significance for Textual Studies." In *Worlds of Words: A Tribute to Arne Zettersten*, edited by Cay Dollerup, 93–107. Oslo: Department of British and American Studies, University of Oslo, 2004.

Burnley, David, and Alison Wiggins, eds. *The Auchinleck Manuscript*. Version 1.1. Edinburgh: National Library of Scotland, 2003 [cited 13 June 2012]. Available from http://www.nls.uk/auchinleck.

Fink, Gernot A. *Markov Models for Pattern Recognition: From Theory to Applications*. Berlin: Springer, 2008.

Hoover, David L. "The Tutor's Story: A Case Study of Mixed Authorship." *English Studies* 93 (2012): 324–39.

McSparran, Frances. "The Language of the English Poems: The Harley Scribe and his Exemplars." In *Studies in the Harley Manuscript: The Scribes, Contents, and Social Contexts of British Library, Harley 2253*, edited by Susanne Fein, 291–426. Kalamazoo, MI: Medieval Institute Publications, 2000.

Pearsall, Derek, and Ian C. Cunningham, eds. *The Auchinleck Manuscript: National Library of Scotland, Advocates' MS. 19.2.1*. London: Scolar Press, 1979.

Robinson, Peter, and Elizabeth Solopova. "Guidelines for Transcription of the Manuscripts of the *Wife of Bath's Prologue*." In *The "Canterbury Tales" Project Occasional Papers*, edited by Norman F. Blake and Peter Robinson, 19–52. Oxford: Office for Humanities Communication, 1993.

Rybicki, Jan. "Alma Cardell Curtin and Jeremiah Curtin: The Translator's Wife's Stylistic Fingerprint." Paper presented at the Digital Humanities Conference, Stanford University, 20 June 2011. Polish-language version published as "Ślady żony tłumacza. Alma Cardell Curtin i Jeremiah Curtin." *Przekładaniec* 24 (2011): 90–110.

Smith, Jeremy J. "The Trinity Gower D-Scribe and his Work on Two Early *Canterbury Tales* Manuscripts." In *The English of Chaucer and his Contemporaries: Essays by M. L. Samuels and J. J. Smith*, edited by Jeremy J. Smith, 51–69. Aberdeen: Aberdeen University Press, 1988.

Stanley, Eric. "Rhymes in English Medieval Verse: From Old English to Middle English." In *Medieval English Studies Presented to George Kane*, edited by Edward D. Kennedy, Ronald Waldron and Joseph S. Wittig, 19–54. Woodbridge: D. S. Brewer, 1988.

Staub, Adrian, and Keith Rayner. "Eye Movements and On-Line Comprehension Processes." In *The Oxford Handbook of Psycholinguistics*, edited by M. Gareth Gaskell, 327–42. Oxford: Oxford University Press, 2007.

Stolcke, Andreas. "SRILM: An Extensible Language Modeling Toolkit." In *Proceedings of the 7th International Conference on Spoken Language Processing*, edited by John Hansen and Bryan Pellom, 901–4. Denver, CO: ISCA, 2002.

Taylor, Andrew. "The Myth of the Minstrel Manuscript." *Speculum* 66 (1991): 43–73.

Traxler, Matthew J., and Morton A. Gernsbacher, eds. *Handbook of Psycholinguistics*. 2d ed. Amsterdam: Elsevier, 2006.

University of Oxford Text Archive. "The Auchinleck Manuscript" [cited 13 June 2012]. Available from http://ota.ahds.ac.uk/headers/2493.xml.

Vinaver, Eugène. "Principles of Textual Emendation." In *Studies in French Language and Literature Presented to Professor Mildred K. Pope*, compiled by Olwen Rhys, 351–69. Manchester: Manchester University Press, 1939.

Witten, Ian, and Timothy Bell. "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression." *Institute for Electrical and Electronics Engineers (IEEE) Transactions on Information Theory* 37 (1991): 1085–94.