Statistical Analysis of Digital Paleographic Data: What Can It Tell Us?

Murray McGillivray
University of Calgary
mmcgilli@ucalgary.ca

Abstract

Manuscript transcription in the Cotton Nero A.x. Project is at a graphetic level and captures each distinguishable glyph used by the scribe. When the transcription is organized as a series of XML entities within a codicological DTD a search-and-count algorithm can be appied to the database of graphetic information. Initial statistical analysis of the data reveals dramatic changes in the scribe's writing system at two points in the manuscript that are roughly coincident with quire boundaries (and also textual boundaries). Hypotheses that will guide further investigation of this phenomenon include the possibility that substantial gaps of time separated the scribe's work in copying the four Middle English poems that make up the manuscript.

KEYWORDS: Statistical analysis, paleography, textual analysis, medieval literature, manuscript, codicology, Pearl-poet, Gawain-poet.

Introduction

The discipline of paleography began from the need to read what the etymology of the term refers to: "old writing." Older handwritten documents were difficult to read because of their unfamiliar handwriting, and, in the case of some ancient, most medieval, and many Early Modern manuscripts, because of abbreviations, brevigraphs, and other unfamiliar deployments of alien sign systems that were no longer in active use. The importance of paleographical study through most of its history has been in the tools that it offers to allow researchers the opportunity for what I will broadly call "transcription," that is, the work of deciphering a text in its location in old documents and placing an equivalent for it in a different sign system within more modern communications technologies, in the first instance, the technology of print. Starting from an almost indecipherable mouldering document, we are able to make a bright clear new document, easily readable because it follows the norm of our own systems of writing and communication. This is not to deny the antiquarian impulse, the genuine desire for contact with the past, in its antiquity and foreignness very specifically, that has motivated almost all modern researches into the musty archives of our cultures, even those researches that also had quite striking political, religious, or cultural agendas relevant to their own times. But the study of paleography arose for reasons of cultural translation, to make the unfamiliar into the familiar, the foreign and antiquated into something acceptable within the frame of the self-same and the modern, and it is taught and studied primarily as an ancillary discipline whose main purpose for most students might be described as "text capture," much as we speak of "image capture" in talking about the various ways in which graphic images can be obtained and processed for use in computer systems.

1. Representation of the Text

As is the case with graphic images, the arrival of the computer has made new things possible in the world of transcription. Not only is it now easily possible to capture an image of a textual document specifically for scholarly use within a computer system, a development that has led to the rise of the "image-based edition," such as the Electronic Beowulf and the Piers Plowman Archive, but various contrivances within computer systems allow a richer representation of the text that is to be captured even within a "text file" than what was possible with pre-digital print technology. Some of those contrivances and technologies for text are proprietary and subject, therefore, to the vagaries of consumer preference and the accidents, in general, that attend upon commercial enterprise, such as bankruptcy and absorption by Microsoft. It is really only appropriate for humanities scholars to investigate and work with those technologies that have more prospect of sufficient longevity to sustain the products of research to reach something like their useful lifespan in the ongoing humanities research conversation. This means choosing those that are not proprietary and are open to all users without dramatic entrance premiums, that are widely implementable across different hardware and operating system configurations, that are widely enough adopted to sustain the development of cheap software tools, and that are flexible enough to do what scholars want. Currently, the complex of technologies developed around HTML and XML is the best choice for all these reasons.

1.1 Transcription of Middle English

Efforts to use the HTML/XML complex of languages, at the time involving primarily SGML (the precursor of XML), for the transcription of early-period primary sources date to the early 1990s. Those efforts were everywhere on the World Wide Web at the time, but appear most thoughtfully in the work of the Text Encoding Initiative, the Canterbury Tales Project as an exemplary implementation avant la letter of TEI recommendations, and Ian Lancashire's Renaissance Electronic Texts (RET) series and its Guidelines. To varying degrees, these pioneering model recommendations and implementations embody a realization that digital transcription or text-capture of the writing (also printing in the case of RET) of the past is more flexible and more able to respond to early sign-systems and graphic communication methods than had been the case with the printed edition. For example, the Canterbury Tales Project editors developed a transcription standard that recorded not just a mapping of the medieval scribes' writing systems to the printable letters available in lower and upper ASCII, at the time a practical limit to what could be done without development of a special font, but a fully "graphemic" rendering of the scribes' work (which did in fact require the development of a special font)

In theory, one can represent anything in the computer. At one extreme one could make a "graphic" representation, in which the limitless repertoire of marks in a manuscript is matched by a limitless repertoire of computer signs. At the other extreme one could make a "regularized" representation, in which the manuscript is transcribed as if for a printed edition, limiting the signs used and regularizing the spelling... Our choice ... lay between a graphemic transcription, aiming to preserve information about distinct spellings in the manuscripts, and a graphetic transcription, aiming to preserve all information about distinct letter forms in the manuscripts. (Robinson and Solopova, 22-24)

The Canterbury Tales Project font, and the SGML entities that underlay it, included the minuscule and majuscule forms of the two Middle English letters thorn and yogh, a total of twelve signs of abbreviation, six superscript letters, four "characters occurring at word ends" and seven medieval punctuation marks in addition to the twenty-five relevant letters (i.e. excluding "z") of the modern alphabet (Robinson and Solopova, 29). This, however, appears a simplification of the symbol-sets of early textuality compared to the recommendation of Lancashire (which admittedly refers to a different transcription situation that perhaps does not have all of the complications described by Robinson and Solopova):

> [A] character is any single unit in a writing system, that is, any unit that on any occasion in a text stands alone in a distinctive way. In English and other languages using the Roman alphabet, these characters include letters, numbers, boundary symbols (such a space, tab, carriage-return/line-feed, punctuation marks, brackets such as parentheses and braces, quotation marks, the apostrophe, accents, and a host of special units like the asterisk) and abbreviations like the ampersand. The Renaissance compositor's typebox might be considered the basic character-set for most RET books, but hand-written scripts such as Anglicana, secretary, bastard, and italic hands offer many additional characters. [...] Renaissance electronic editions cannot assume anything about the functions or symbolic roles of the characters in [the] character set [of the Renaissance]. The tagging system should uniquely identify characters while postponing demands that they be declared to take on any one role. (Lancashire Guidelines, "4. The Renaissance Character Set")

On the whole, this more stringent position, essentially a recommendation for transcription of all of the individual textual signs or "graphs" used in a document, is more appropriate to the real textuality of Middle English manuscripts than the position finally taken by Robinson and Solopova (to be fair, it is equivalent to a "level" of transcription for which they evince some sympathy, originally had experimented with, and were deterred from primarily by the practicalities of their immense project [Robinson and Solopova 25). 1.2

1.2 Cotton Nero A.x. project

Lancashire's procedure of representing all of the distinct graphic signs of the original document in the transcription rather than replacing them with a more restricted character set inspired by modern typography, a procedure that medievalists would call "graphetic transcription," is the basis for the general approach we have adopted in the Cotton Nero A.x. Project, a team project to produce a digital edition of British Library MS Cotton Nero A.x. (Art. 3), the famous Middle English manuscript uniquely containing the poems Sir Gawain and the Green Knight, Pearl, Patience, and Cleanness, four of the most-studied poems in the canon of Middle English literature. This manuscript was written in the late fourteenth or early fifteenth centuries by a single scribe in a single script, which does simplify the task of transcription. We have transcribed with a separate digital representation each separate graphic symbol used by the scribe of the manuscript, with special provisions for graphic symbols that are used with multiple phonetic meaning (a group that were a bugbear for Robinson and Solopova and one of the factors that inclined them in the direction of "graphemic" transcription). This gives the Cotton Nero A.x. Project a "character set" of about a hundred signs in addition to alphabetical characters from the modern alphabet, each of the extra-alphabetical signs having a distinct representation as an XML entity in our transcription scheme.

In addition to the minuscule and majuscule letters of the late medieval English alphabet, the transcription scheme thus includes XML entities to act as separate transcriptions for variant letter forms (such as short and long "s"), to represent abbreviations and brevigraphs, and to distinguish junctures (that is, letters written together as one graphic unit, rather like print-era ligatures) from the separately-written letters. Both the scheme itself, with photographs of the "characters" that are included in the scheme together with their XML transcription equivalents, and the full manuscript transcription together with an XML DTD (Document Type Definition) and a CSS (Cascading Style Sheet) stylesheet for display, are currently available on the project Web siteurl. The site also includes a fuller description of the project and its goals and participants.

There is certainly an additional cost to "graphetic" transcription compared to the looser representation offered by Robinson and Solopova's "graphemic" transcription and adopted from them by other projects (or compared to the representations offered in analogous print-form diplomatic transcriptions such as the one in the Ruggiers Hengwrt facsimile). The cost is not necessarily in correctness, as Robinson and Solopova imply that it must inevitably be (25) --"accuracy," their word for correctness, is a word to be avoided in this context since it might refer either to the appropriateness of the transcription scheme itself to the materials being transcribed (how fully the writing system of the scribe is represented) or to how correctly the scheme is implemented (how error-prone the transcribers and proofreaders are) -- but in the additional time required to develop an appropriate transcription scheme, in the additional training time required to get student research assistants to reach an adequate standard of transcription acuity, and in the additional time required for seasoned scholars to review and correct the work of the initial transcribers. However, some of these additional costs might disappear were it possible to implement an image-recognition-based automatic transcription program of the kind that might seem to be forecast by the development of Blobworld and other attempts to sort through large image collections automatically using fuzzy methodologies (see also the work of Arianna Ciula) -- Optical Character Recognition is currently too high-strung a process to use on most handwritten documents, as anyone who has tried scribbling into a PalmPilot can attest.

It is therefore appropriate to ask whether the digital representation of paleographic data that is contained in a full graphetic transcription is a goal worth pursuing in the long term for editors of medieval and Renaissance manuscripts in general. The Cotton Nero A.x. Project transcription might be better described as an experiment in itself in graphetic transcription than a "proof of concept": the goal has never been to show that there are practical benefits to close representation of the scribe's writing system in the transcript, because there are enough intrinsic benefits in the case of this valued manuscript to make such transcription worthwhile in itself. However, the completion of the transcript in close-to-final draft does give us a lot of paleographic data to look at, particularly with computer tools.

2. Analysis of the Distribution of Paleographic Data

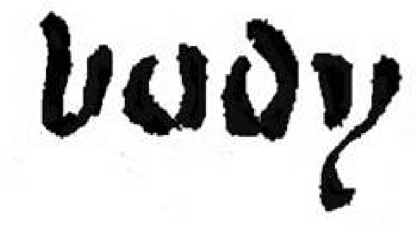

Figure 1

One of the first things that I wanted to look at when the transcription was finished and proofread to a decent level of accuracy was the distribution of junctures (also called "biting" by paleographer -- the term "juncture" is used by Samuel Harrison Thomson and is a more accurate description of what the scribe actually does). Junctures, as stated above, are an equivalent in medieval handwriting of the printed ligature. Instead of writing two separate letters, the scribe writes a kind of monogram combining them. For example, the letters "b" and "o" may be written separately (figure 1) or they may be in juncture (figure 2) : the left-hand stroke of the letter "o" is curved in such a way as also to do duty as the right stroke that would form the compartment of a separated letter "b." In the script used in MS Cotton

Nero A.x., which derives from a rounded Gothic or textura book hand but has features from Anglicana, juncture is common (as it is in Gothic hands) between consonants that have right-facing bows, such as "b" "h," and "d," and vowels following them that have curved left sides, so in this hand, "a" "e," and "o." Research assistant Samantha Barling had suggested to me as she was working on transcription that the use of juncture between "b" and following vowel "o" seemed to drop off towards the end of the manuscript, and this was an oddity I thought worth investigating. One would think that the habitual formation of junctures between letters would be a near-automatic part of the scribe's writing system that would be unlikely to change in any substantial way in the course of copying out a short manuscript like Cotton Nero A.x. (some 90 folios of 36 lines per side, so not an overwhelming copying job).

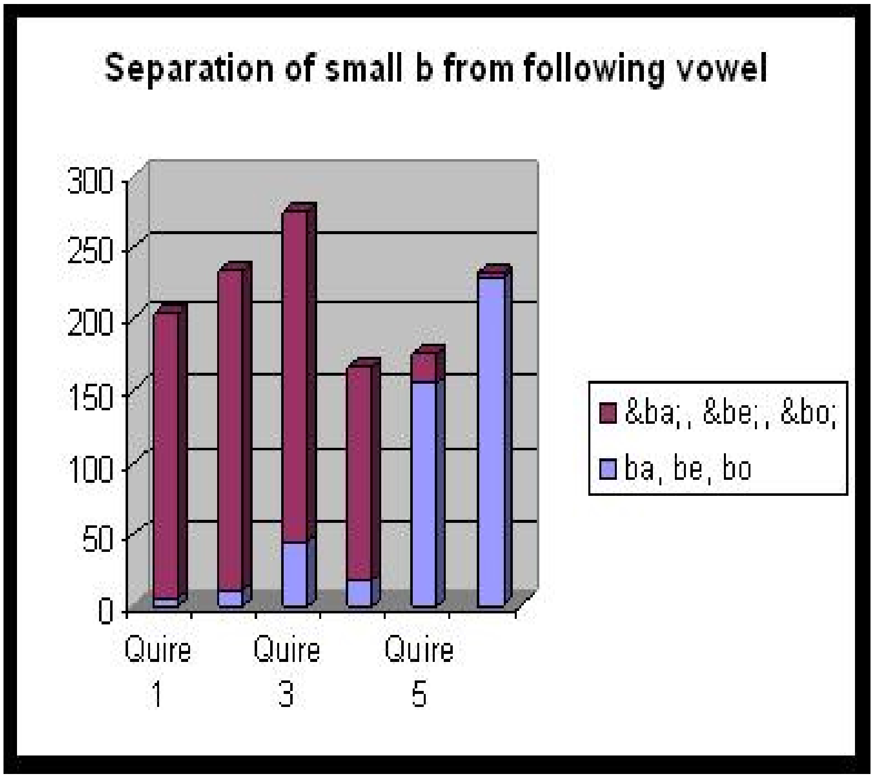

Figure 2

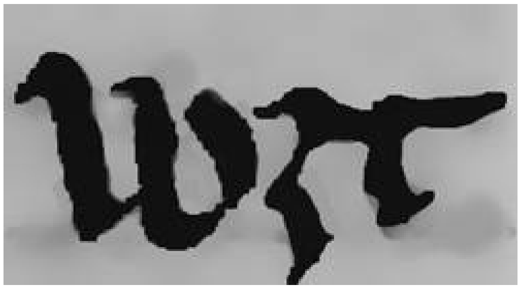

Figure 3

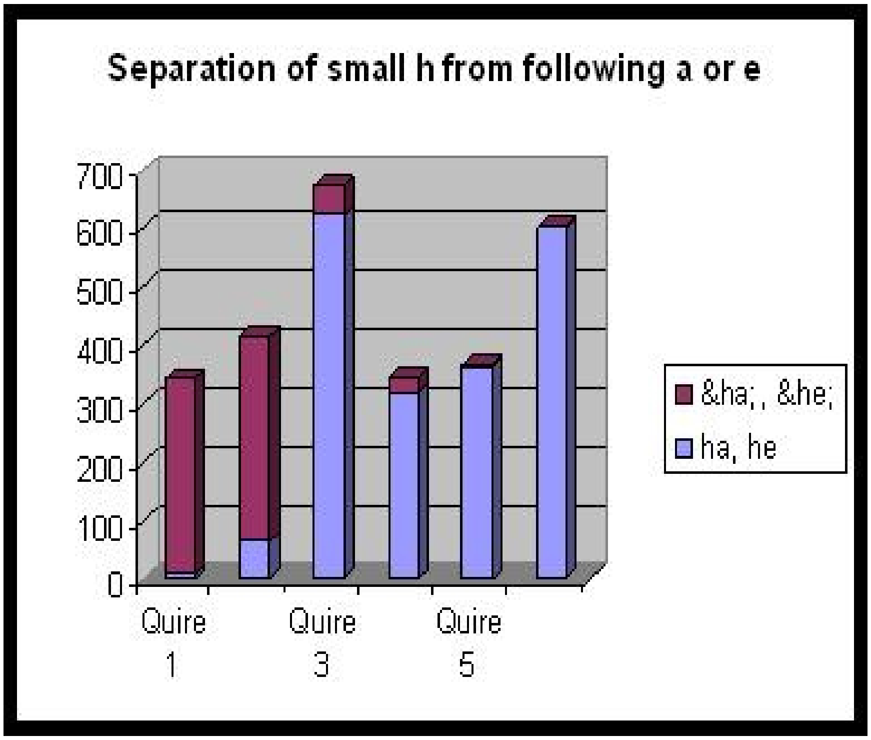

Figure 4

Using code already on hand from a short program I had written to locate instances of particular XML entities as part of our proofreading process for the transcription, I developed a piece of Java that works with a SAX ("Simple API for XML") parser to count instances of letters and other symbols, whether represented in the transcription with the letters of the modern alphabet or with XML entities, within particular divisions of the manuscript. The routine, called "analyzeThis," was run using quires, the gatherings of parchment that make up the structure of a medieval handwritten book and that therefore form one of our basic XML units ("elements") for the transcription, as the chosen level of detail. Results are presented in graphic form for the juncture of "b" with following "a" "e", or "o" versus the letters written beside one another in a word but not joined (figure 3), the juncture of "h" with following "a" or "e" versus the separated letters (figure 4), and the juncture of "d" with "e" versus the separated letters (figure 5).

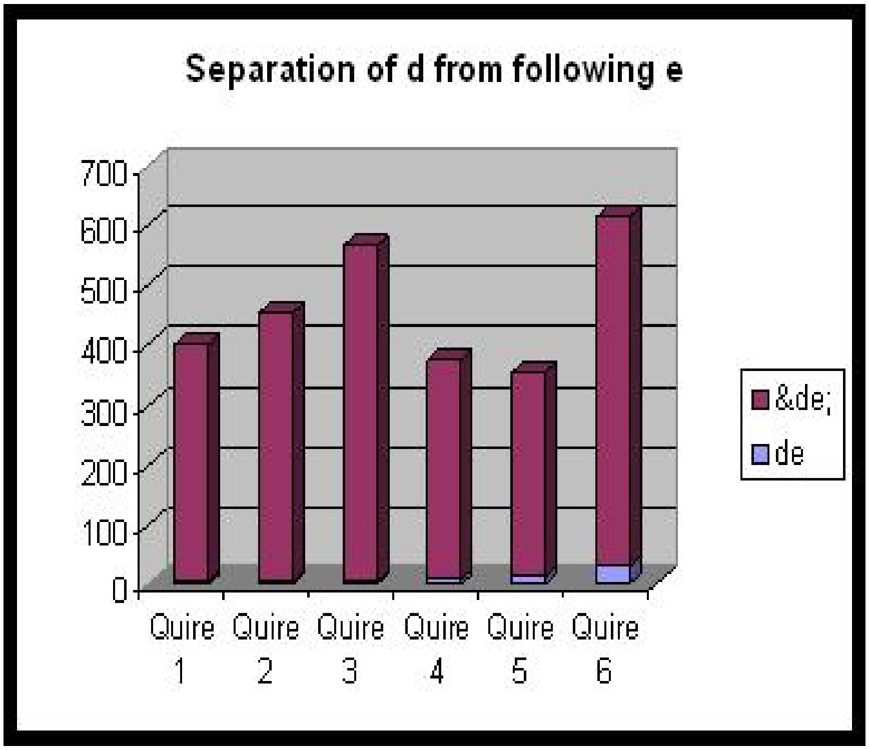

Figure 5

This initial run through the data certainly confirms Ms. Barling's impression as a transcriber. There is a dramatic drop-off in the use of a juncture for "ba," "be," and "bo" (represented in the graph labels of figure 3 by the XML juncture entities, "&ba;" "&be;" "&bo;", that were adopted by the project), a drop-off that occurs apparently quite suddenly somewhere around the border between quire 4 (folios 75 to 86) and quire 5 (folios 87 to 98) of the manuscript. In the first quire, the scribe's use of juncture approaches 100% of the cases where the round-sided vowels follow letter "b"; by the sixth quire, use of juncture in these situations is near 0% of the cases. In quire 4, on the order of 90% of the cases exhibit juncture; in quire 5, almost exactly the reverse is the case.

Figure 5A similar pattern, but around a different break point, occurs with respect to "h" followed by "a" or "e" (see figure 4); in this case, the situation where there would be juncture of "h" and "o" is too rare to contribute statistical value and seems anyway not to participate in the pattern). Again, there is almost complete adherence to the practice of writing the juncture at the beginning of the manuscript: fewer than 10 of the more than 300 instances of "h" followed by "a" or "e" are written with the separate letters in quire 1. And there is almost complete abandonment of the practice at the end of the manuscript: more than 98% of some 600 cases in quire 6 are separate letters rather than junctures. Again, there appears to be a dramatic drop-off, where quire 2 (folios 51 to 62) has a very high proportion of junctures and the following quire 3 (folios 63 to 74) has a very low proportion.

For contrast, the expected kind of pattern is shown for "d" followed by "e" (see figure 5.) Here the scribe is relatively constant in his joining of the two letters, with perhaps a slight upward trend in instances of the separation observable in the last two columns of the graph. This behaviour does not need explaining. The scribe simply has his own habitual way of writing this sequence of letters, a sequence that typically occurs at word-end but also occurs in other environments, in which he adds a hook and an upward angular stroke to the right side of the "d" in order to complete an attached "e." That there are a few more instances of the two letters in succession written separately towards the end of the manuscript is worthy of note, but seems to be a fact of the same order as the gradual degeneration of the calligraphic quality of a hand in the course of writing a book noted by Malcolm Parkes (21; Plate 21).

What do the other two, more dramatic, shifts in scribal behaviour, one between quires 2 and 3 and the other between quires 4 and 5, indicate? Just as it would help to situate Parkes's instance of what appears to be degeneration of a hand in the course of writing a book to have an array of examples of books from single scribes for comparison (is it usual or unusual for a scribe to start a manuscript using more carefully-formed letters than he uses as he finishes it?), it would be useful in assessing the statistical anomalies in Cotton Nero A.x. to have an array of data with which they could be compared. From this point of view, it is to be regretted that Angus McIntosh's call to give to written Middle English more serious study as "a system operating in its own right -- a system specifically of written language" ("Analysis" 53), a call first made in 1956 in "The Analysis of Written Middle English" and repeated more pointedly in 1974 in "Towards an Inventory of Middle English Scribes" and in 1975 in "Scribal Profiles from Middle English Texts", has been one that the very large projects of early adopters of computer technology for Middle English in the 1990s have been unable to heed in their transcription policies. In many ways it would have been better for Middle English studies to have entered the age of digital editions with a series of smaller projects in which basic principles could have been tested and established.

2.1 Hypotheses

In the absence of the comparative information that might have been produced if such smaller projects had given to individual manuscripts the kind of detailed attention that we have given to Cotton Nero A.x., it is only possible to formulate uncertain hypotheses:

> 1. Although the computer analysis described above proceeded according to the codicological units of manuscript construction, there does not seem to be correlative evidence in the construction of the manuscript that would suggest that any fact of that construction is in itself relevant.
>
> 2. The boundaries between quire 2 and quire 3, and between quire 4 and quire 5, however, are also close to textual divisions in the manuscript, between Pearl and Cleanness in the first case and between Cleanness, Patience and Sir Gawain in the second (Patience extends across the quire boundary and is short enough that its situation with respect to the abrupt change in writing system cannot be established with the rough statistical methodology described in this paper). These textual divisions may be relevant.

> 3. The changes are small enough that intentional alteration of the writing system by the scribe seems unlikely since too unmotivated—and the possibility that one or other of the changes was done on purpose is argued against by the fact that the other change did not happen at the same time although it is similar.
>
> 4. It is a reasonable hypothesis, then, that the changes took place unconsciously, and that they therefore indicate the intervention of some event or process that altered the scribe's approach to the writing of his script without his knowledge.
>
> 5. One possibility worth considering is that significant time intervened between the copying of Pearl and that of Cleanness and then again either after the writing of Cleanness and before Patience or between Patience and Sir Gawain. Especially if such an interval of time were occupied with writing in a script that did not use junctures, the scribe might easily forget where they were to be formed in returning to the Cotton Nero A.x. script.

The last step in this chain of hypotheses is a possibility that might have relevance for a variety of questions, including questions about the nature of the exemplar or exemplars used by the scribe, and extending even to the question of the sequence of composition of the four poems. Certainly the entire sequence of hypotheses needs to be subjected to the kind of scrutiny it will only get when several transcriptions of different manuscripts done at the graphetic level have been made available to scholars, but it is also worth observing that part of its uncertainty is due to the coarseness of the tools -- we cannot even say positively at this point whether the changes in scribal behaviour are more closely associated with quire boundaries or the boundaries between works. An obvious further step is to refine the computer tool in such a way as to be able to see exactly where the change in scribal behaviour takes place, if indeed there is a place that can be located more precisely.

Conclusion

McIntosh argued that graphetic analysis could be combined with linguistic analysis to create profiles of individual scribes, which he hoped would both illuminate aspects of the nature of written Middle English and, eventually, contribute also to our understanding of spoken Middle English. The evidence presented in this paper suggests that there could well be wider uses still for paleographic data collected in the process of transcription of Middle English primary sources, including, but surely not limited to, the study of scribal behaviour and analysis of the interrelationship between the writing system used by a scribe and the process of production of the manuscript. Collecting and analyzing this data might give us a better understanding of the linguistic processes, including especially the graphic processes of written language, that attended on the production of individual manuscripts or groups of manuscripts, a matter of high interest for linguists and literary scholars who perforce deal with written texts. Paleographic study of medieval texts, in conjunction with contemporary methods of computer analysis, can become one of the new and promising tools we can use to understand the written production of the past. It might seem that transcription into the machine codes of the computer is yet another way of making medieval texts conform to the standards of the modern age, but paradoxically we have never been able to be closer to the writing systems of the past. Statistical and other analyses of digital paleographical data garnered through careful study of medieval documents can provide a new window into the nature of medieval textuality.

Works Cited

Blobworld: Image Retrieval Using Regions. <http://elib.cs.berkeley.edu/photos/blobworld/>.

The Canterbury Tales Project. Director Peter Robinson. <http://www.cta.dmu.ac.uk/projects/ctp/>.

Ciula, Arianna. (2003). "Computacional [sic] Suggestions to [sic] Paleographical Analysis." *lamusa digital 3*. <http://www.uclm.es/lamusa/>.

The Cotton Nero A.x. Project. General ed., Murray McGillivray. <http://www.pearl-ms.ca/>

The Electronic Beowulf. (1999). Ed. Kevin Kiernan, with Andrew Prescott, Elizabeth Solopova, David French, Linda Cantara, Michael Ellis and Cheng Jiun Yuan. London: The British Library.

Lancashire, Ian. *Renaissance Electronic Texts: Encoding Guidelines*. (1994). Toronto: Centre for Computing in the Humanities, University of Toronto. <http://www.library.utoronto.ca/utel/ret/guidelines/guidelines0.html>.

McIntosh, Angus. (1956). "The Analysis of Written Middle English." *Transactions of the Philological Society*: 26 – 55.

---. (1974). "Towards an Inventory of Middle English Scribes." *Neuphilologische Mitteilungen* 75: 602 – 624.

---. (1975). "Scribal Profiles from Middle English Texts." *Neuphilologische Mitteilungen* 76: 218 – 235.

Parkes, M.B. (1980). *English Cursive Book Hands*, 1250 – 1500.London: Oxford UP, 1969. Reprint Berkeley: U California P.

The Piers Plowman Electronic Archive. Project Director Hoyt N. Duggan. http://www.iath.virginia.edu/seenet/piers/

Renaissance Electronic Texts. (1997). General ed., Ian Lancashire. Toronto: Web Development Group, University of Toronto Library. http://www.library.utoronto.ca/utel/ret/ret.html

Robinson, Peter, and Elizabeth Solopova. (1993). "Guidelines for the Transcription of the Manuscripts of the Wife of Bath's Prologue." The Canterbury Tales *Project Occasional Papers* 1. Oxford: Office for Humanities Communication: 19 – 52.

Ruggiers, P.G. (1979). *A Facsimile and Transcription of the Hengwrt Manuscript, with Variants from the Ellesmere Manuscript*. Variorum Chaucer. Norman, Oklahoma: U Oklahoma P.

The Text Encoding Initiative. <http://www.tei-c.org/>

Thomson, Samuel Harrison. (1969). Latin Bookhands of the Later Middle Ages, 1100 – 1500. London: Cambridge UP.