LEXOMIC ANALYSIS OF ANGLO-SAXON PROSE: ESTABLISHING CONTROLS WITH THE *OLD ENGLISH PENITENTIAL* AND THE OLD ENGLISH TRANSLATION OF OROSIUS

Abstract: In this paper we demonstrate that "lexomic" techniques of computer-assisted statistical analysis, originally validated for Old English poetry, can be adapted and applied to Anglo-Saxon prose texts. The methods we describe employ hierarchical agglomerative cluster analysis to find patterns of vocabulary distribution. These patterns, represented visually as tree diagrams, or *dendrograms*, can indicate the source structure or the affinities of Old English texts. Comparing the dendrogram geometry of multiple editions of the *Old English Penitential* allows us to determine that the methods can produce consistent results even for critical editions made from the collation of multiple manuscripts. Analysis of the Old English translation of Orosius's *Historia* demonstrates that the techniques can detect where an author has used for a given section of his text sources different from those of the main body of the text. We conclude that lexomic methods are a useful new tool for the analysis of Old English prose. **Keywords**: Lexomics, computer-assisted analysis, digital humanities, penitentials, Orosius, *Historiarum adversus paganos libri septem*, Alfredian translations, sources, editions.

Resumen: En este artículo demostramos que las técnicas lexómicas de análisis estadístico asistido por ordenador, válidas originalmente para la poesía en inglés antiguo, pueden adaptarse y aplicarse a textos anglosajones en prosa. Los métodos descritos emplean análisis jerárquicos de clústeres aglomerativos para encontrar patrones en la distribución del vocabulario. Tales patrones, representados visualmente mediante diagramas arbóreos o *dendogramas*, pueden revelar la estructura de la fuente o las afinidades de textos en inglés antiguo. Comparar la geometría del dendrograma de ediciones múltiples del *Old English Penitential* permite determinar que esos métodos pueden producir resultados consistentes incluso con ediciones críticas hechas mediante la colación de múltiples manuscritos. El análisis de la traducción anglosajona de la *Historia* de Orosio demuestra que las técnicas pueden detectar dónde un autor usó para una sección fuenrtes distintas de las del cuerpo principal de texto. Concluimos que los métodos lexómicos son instrumentos útiles para el análisis de la prosa anglosajona. **Palabras clave**: lexómica, análisis asistido por ordenador, humanidades digitales, penitenciales, Orosio, *Historiarum adversus paganos libri septem*, traducciones alfredianas, fuentes, ediciones.

IN A RECENT SERIES OF PAPERS OUR RESEARCH GROUP HAS demonstrated the value of combining computer-assisted, statistical analysis with traditional, philological approaches to medieval texts. This *lexomic*[1] approach, which detects

[1] Coined by Betsey Dexter Dyer in 2002, the term "lexomics" is derived by analogy from "genomics" (Dyer 2002) and first appeared in *Genome Technology* 1.27 (2002).

ISSN: 1132–631X Phoebe Boyd, Michael D. C. Drout, Namiko Hitotsubachi, Michael J. Kahn, Mark D. LeBlanc & Leah Smith, *SELIM* 19 (2012): 7–58

patterns of vocabulary distribution that are not otherwise visible to the unaided eye,[2] has already shed new light on poems in Anglo-Saxon and on medieval Latin prose and poetic texts,[3] and methods originally developed for the analysis of Old English poetry can, we believe, be adapted to investigate texts from the much larger corpus of Anglo-Saxon prose. In this paper, therefore, we use lexomic methods to analyze the Old English penitentials and the Anglo-Saxon translation of Orosius's *Historiarum adversum paganos libri septem,* demonstrating not only the utility of the methods but the specific ways they must be modified in order to be applied to prose texts, which present a particular suite of problems. Although the challenges presented by text length, manuscript variation and editorial practice are substantial, lexomic analysis of Anglo-Saxon prose provides a new channel of information that can both support conjectures made by previous scholars and also open up new lines of inquiry.

1 Lexomic Methods

Lexomic methods blend techniques from bioinformatics,[4]

[2] The development of some of the lexomic methods discussed in this chapter were supported by the National Endowment for the Humanities, which sponsored the research with two grants, NEH HD-50300-08, *Pattern Recognition through Computational Stylistics: Old English and Beyond*, and NEH PR-50112011, *Lexomic Tools and Methods for Textual Analysis: Providing Deep Access to Digitized Texts*. Any views, findings, conclusions, or recommendations expressed in this article do not necessarily reflect those of the National Endowment for the Humanities.

[3] Forthcoming papers demonstrate that lexomic methods can also be used to analyze texts in Old Norse, 20th-century Modern English (both drama and prose) and 17th-century English (drama).

[4] Bioinformatics treats nucleobases in DNA as an alphabet, combinations of nucleobases as "words," and genomes as texts. In their analyses, bioinformaticists have re-invented a number of techniques originally developed by philologists, such as the tracing of descent through shared error. See, for example Dyer *et al*. 2007.

computational stylometry,[5] and traditional textual analysis (including philology, source study, historical contextualization and close reading). Using the high-quality electronic editions of medieval texts now available to researchers, we employ computer-assisted statistical techniques to identify patterns, which we then interpret using traditional literary methods. At the beginning of our research, the computational methods told us where in a text to look, while the traditional methods explained what our findings meant, but as our research has progressed we have found that this expected pattern has at times been reversed, and our methods have evolved into a series of *iterate and test* processes that integrate all the tools at our disposal.

Lexomic methods differ slightly from pioneering stylometric analyses in two major ways. First, although most researchers analyze subsets of words in a text (function words or content words, for example), we include every word in our analyses. Second, while computational stylometry has traditionally focused on whole works, we divide our texts into segments and analyze the relationships of these to each other. Also, although the information we recover with our methods may have some bearing on questions of authorship, our analyses have to this point do not focused primarily on author identification but instead on a text's sources or affinities.[6]

The techniques discussed here all can be performed using our software, which is browser-interfaced and freely available in the

[5] Pioneers of computational stylometry include John Burrows and David Hoover. Burrows 2003 uses statistical analysis of "function words" (prepositions, conjunctions, pronouns) to create textual "signatures" for various writers, which he then uses to attribute authorship in a set of English Restoration poems. Hoover 2004 further refined Burrows's methods and applied them to prose in third-person American novels.

[6] Admittedly, sources and affinities can have some bearing on authorship, and we have used lexomic methods to support the case for identifying *Guthlac B* as being written by Cynewulf (Drout *et al.* 2011: 323–326).

Lexos Integrated Workflow at http://lexos.wheatoncollege.edu.[7] We begin by *scrubbing* an electronic edition of a text, removing punctuation, changing capital letters to lower-case, and deleting formatting codes and other tags.[8] Scrubbing allows us to compare like to like, making certain that we count *king* as being the same word as *King* and (*king*) and not counting commas or periods as "words." After the text is scrubbed we divide it into segments and then tabulate the words in both the entire text and in each segment.[9] In order to allow us to compare segments of different sizes, we compute relative frequencies for each word by dividing the number of times the word appears in a segment by the total number of words in that segment.[10] From this data we produce an *n*-dimensional array for each segment, where *n* represents the number of distinct words used in the entire collection of texts being studied.[11]

[7] Documentation and instructional videos and web pages are available at http://wheatoncollege.edu/lexomics/introduction-lexomics. The research for this paper was performed using a previous iteration of the tools, which are preserved in the Lexomics Tool Archive: http://wheatoncollege.edu/lexomics/tool-archive.

[8] The program *Scrubber,* written primarily by Richard Neal, was used for these purposes. It can also be used to lemmatize a text or to modify special characters. *Scrubber* is now a part of the Lexos Integrated Workflow. The version of *Scrubber* used to perform the research in this paper is preserved in the Lexomics Tool Archive.

[9] The program *DiviText*, written primarily by Amos Jones, was used to cut texts into segments and count the words in those segments. *DiviText* itself is not part of the Lexos integrated workflow, although Lexos provides much of *DiviText*'s functionality. *DiviText* remains accessible in the Lexomics Tool Archive.

[10] If there are 1000 words in a segment and *ond* appears 50 times, we record 50/1000 = 0.05 as the relative frequency of *ond*. If a word appears somewhere in the complete text but not in a particular segment we record 0/1000 = 0 for the word's relative frequency in that segment.

[11] Technically, the scripts use a hash table of arrays. Interested readers are directed to the documented software for specifics.

We then use the free implementation of hierarchical, agglomerative cluster analysis (Mardia *et al.* 1980) within the statistical software package, R (R Development Core Team 2009), to group the segments.[12] This clustering method uses a dissimilarity metric for the grouping of texts without pre-specifying a number of groups. The dissimilarity (or distance) measure is computed for each pair of segments, and these distances are then used to create groupings, or *clades*,[13] of texts by clustering texts that are most similar (i.e., have the shortest distance between them).[14] In the analyses presented in this paper, we employ the most commonly used metric, Euclidean distance,[15] to calculate the distance between the multidimensional averages of the two clades. We then use hierarchical agglomerative clustering to order these distances and construct a branching diagram, or *dendrogram*,[16] of their relationships. The dissimilarity between clades is represented by the

[12] An explanation of the statistical methods, aimed towards humanistic researchers, can be found in Drout 2013: 51–56.

[13] The terminology is borrowed from evolutionary biology (Hennig 1966).

[14] To compare four segments we list all the words in each segment and calculate the relative frequency of each word in each segment. We then compute (4×3)/2=6 distances, one for each pair of segments, calculate the difference between the proportion of a word's use in each segment, square the differences, and total the squared differences from each word. The distance, then, is the square-root of the squared distance.

[15] This metric makes use of all *n* words in a collection of texts to measure the dissimilarity between two texts. We also experimented with Manhattan and Canberra metrics but found no significant difference in the final clustering results. Our software allows researchers to choose among these metrics and between different linkage methods.

[16] The program which creates dendrograms, *TreeView*, was written primarily by Alicia Herbert. *TreeView* is now a part of the Lexos Integrated Workflow. The version of *Tree-View* used to perform the research in this paper is preserved in the Lexomics Tool Archive.

vertical length of the line connecting them.[17] Figure 1 illustrates the similarities of four hypothetical segments or texts. Any level of the branching diagram can be identified as a clade, and we label clades from left to right using Greek letters, first marking all clades at the same level of the hierarchy and then descending to the next level and again labeling left to right. Thus in Figure 1, clade α contains segment A, clade β contains B, C, and D, and clade γ contains only segments C and D. A clade with no subsidiary branches, like clade α, is said to be *simplicifolious*.

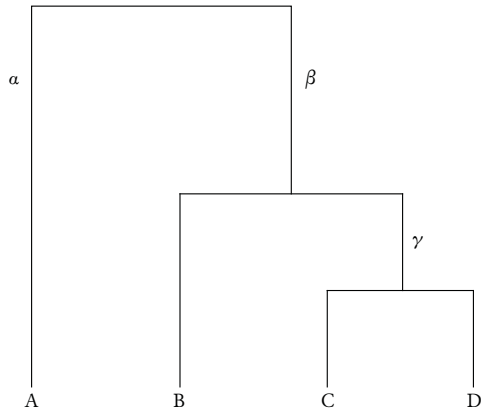

Figure 1. *Sample Dendrogram*

The geometry of the dendrogram indicates that segments C and D in Figure 1 are most similar, segment B is closer to clade γ, which contains both C and D, and segment A is least like the other texts. The vertical distance between segments C and D is smaller than that between the simplicifolious clade α and clade β, indicating that segment A is quite different from the other segments.

[17] In our lexomic analyses the number of words is quite large, so it is difficult for the distributions of any single word to make two segments highly similar or dissimilar. Instead, it takes a great deal of commonality (or difference) in the proportionate use of a wide array of words to create large similarity (or distance) between two texts. See the discussion in Drout *et al*. 2011: 311–315.

Our previous work with Latin poetry and prose and Old English poems has shown that the geometry of a dendrogram can be influenced by the affinities or sources of the texts being analyzed: similar segments or texts tend to cluster together. For example, the Old English poem *Azarias* is paired in a dendrogram with the section of *Daniel* that is known to be very similar to it (both have a recent common textual ancestor; Drout *et al*. 2011: 307–311). In addition, texts with multiple sources produce dendrograms in which the segments are grouped by source: a dendrogram of the Old English *Genesis* places *Genesis B* in a high-level clade entirely separate from *Genesis A,* and the segments of *Daniel* that are based on Latin canticles are separated from the rest of that poem (which is based on the Bible; Drout *et al*. 2011: 326–335). Dendrograms of Latin texts likewise reflect both sources and affinities. The source structure of Alan of Lille's *De planctu naturae* is evident in its dendrogram, as is that of Geoffrey of Monmouth's *Vita Merlini*. Every papal letter quoted in Bede's *Ecclesiastical History* separates from Bede's main text. A dendrogram of the *Gesta Friderici Imperatoris* places chapters by its two authors (Otto of Friesing and his secretary, Rahewin) in separate clades (Downey *et al*. 2012). However, the two types influences—of sources and of affinity—can also conflict with or complicate each other. For example, the segments of the Old English poem *Juliana* in which Cynewulf closely follows the Latin *Vita* that was his source cluster separately from the rest of Cynewulf's corpus (Drout *et al*. 2011: 333–335), and the segment of *Guthlac B* dependent upon the "cup of death" motif (which is not found in Felix's *Vita s. Guthlaci*) appears separately from the rest of that poem and of the signed poems of Cynewulf (Downey *et al*. 2012). In these and similar cases it is essential to use non-lexomic knowledge about the text to interpret the lexomic results rather than relying solely upon dendrogram geometry.

Although the results of lexomic analysis of texts whose sources, affinities or structures are well understood are not counter-intuitive or even surprising, they are nevertheless quite valuable.

Even if a particular lexomic analysis tells us nothing entirely new about a text, the correlation of dendrogram geometry with previously existing knowledge gives us some confidence in analyses of texts whose authorship, sourcing or structures are unknown or controversial. The dendrograms of the known texts serve as *controls* for the dendrograms of the unknowns; if the former are consistent, we are not unreasonable in trusting the latter. But in order to establish such controls, it is important to determine which variables of orthography, manuscript variation and editorial practice are significant, so that we can compare like to like.

2 Corpus-Specific Parameters

Thanks to initiatives both organizational and individual, a significant number of medieval texts are now available in electronic form. Most important for our purposes is the complete corpus of Anglo-Saxon assembled by the *Dictionary of Old English*.[18] But although the *DOE Corpus* contains a high-quality edition of every known Anglo-Saxon text in a well-curated archive, we still must address some corpus-specific questions before we can perform lexomic analysis.

First, there is the problem of orthographic variation. Because our software compares and counts words according to exact identity, orthographic variation has the potential to obscure significant patterns or to create statistical artifacts in our analysis. We must therefore process the texts in such a way to eliminate trivial variation without losing significant data. This processing must be customized to each writing system. For the Old English corpus the most significant orthographic variations are between *thorn* ⟨þ⟩ and *eth* ⟨ð⟩—both of which are used to represent voiced

[18] The *Dictionary of Old English* can be accessed at http://www.doe.utoronto.ca/index.html; a subscription is required. The tools on the lexomics.wheatoncollege.edu website produce data about the corpus but do not distribute the corpus as a whole.

and unvoiced interdental fricatives—and among the Tironian note ⟨⁊⟩, *and* and *ond*.

Scholars have long noted that the distribution of *thorn* and *eth* in the Old English corpus is not phonetically consistent. Unlike Icelandic orthography, in which ⟨ð⟩ generally represents the voiced and ⟨þ⟩ the unvoiced interdental fricative, in Anglo-Saxon either letter can be used represent either sound. The distribution of forms, however, is not entirely random. Some early manuscripts use only ⟨ð⟩, while in later manuscripts the forms are more evenly distributed (Roberts 2006: 20–28), and different scribes appear to have different tendencies to use each symbol, some, for example, seeming to avoid the use of medial *thorn* or initial *eth* but others not following these practices (Klaeber 2008: xxix–xxx, cliv–clvii). David Megginson has shown that there is significant variation in the ratio of *thorn* to *eth* from manuscript to manuscript and from scribe to scribe. He also notes that certain words are consistently spelled with one letter or the other regardless of the phonetic value in the particular context, suggesting, he argues, that the spellings were memorized rather than phonetic.[19] Recent work by our research group shows that substantial variations in the *thorn* to *eth* ratio *within* a given scribe's performance in a given manuscript may be diagnostic of differences in textual source (Chauvet and Drout forthcoming). This variation and its possible significance thus creates two problems. If we treat the variation between *thorn* and *eth* as significant and count *þis* and *ðis* as two distinct words, we may be unable to compare texts that are found in separate manuscripts, since scribal performance might overshadow other kinds of variation.[20] But if we collapse the variation and treat all

[19] Megginson 1993: 35–36, 49–51, 60–62, 100–107 and *passim* in discussions of words that contain ⟨þ⟩ or ⟨ð⟩.

[20] Variation between ⟨i⟩ and ⟨y⟩, which O'Donnell 2005 has shown to be the most common variation in the poetic corpus, has not to this point been a significant problem (with the exception of *Beowulf*).

interdental fricatives as the same, counting *þis* and *ðis* as the same word, we might lose relevant data.

Our solution to these problems is both empirical and mathematical. Our software allows us to *consolidate* texts, by converting all *eths* to *thorns* (or vice versa). We can therefore easily compare dendrograms of the consolidated with those of the unconsolidated texts. To this point, comparing hundreds of dendrograms, we have only found one complete text (*Beowulf*) and two segments (both in *Christ III*) whose locations in a dendrogram change when the texts are consolidated.[21] Research into the characteristics of those two segments is ongoing, but we can conclude that in the vast majority of cases, orthographic variation of *thorn* and *eth* does not affect dendrogram geometry. Mathematically this lack of effect can be explained by the interchangeable nature of the two letters. Even when we count *þis* and *ðis* separately, if the variants are equally distributed in the segments, the distances between texts will not be affected, since the relative frequency of either *þis* or *ðis* will simply be (*þis* + *ðis*)/2 split among the two segments.[22] Only significant concentrations of either orthographic form would affect the dendrogram geometry, and these concentrations appear to be relatively rare in Old English poetry. Furthermore, since the analysis presented in this paper is lexico-morphologic rather than orthographic, we can be reasonably comfortable in using consolidated forms. Nevertheless, we have performed all the experiments discussed here using both consolidated and unconsolidated forms, and the results have been the same.

Anglo-Saxon scribes' use of Tironian note creates a slightly different problem because the grapheme could in Old English represent either *and* or *ond*. Expanding the note to either all *and*

[21] Drout *et al*. forthcoming and Chauvet & Drout forthcoming.

[22] If there were 8 instances of the consolidated word in text A and 6 in text B, the distance between the two texts would be 2. If *thorn* and *eth* are equally distributed, there would then be 4 instances of each orthographic form in text A and 3 in text B. The distance would then still be 2: (4-3)+(4-3)=8-6.

or all *ond*, therefore, has the potential both to obscure existing patterns or to create artifacts, since we cannot know what form the scribe was abbreviating.[23] We could choose not to expand the note, but by so doing we would be privileging the manuscript form of a text over its linguistic expression—a procedure which might at times be useful, but which is not necessarily always justified. Furthermore, because *and*/*ond*/⁊ is the most common word in the Old English corpus, variations in its form affect the geometry of dendrograms in a way that variation between *thorn* and *eth* do not.[24] We therefore use our software to *lemmatize* ⁊, *and* and *ond* to a single form (arbitrarily, *and*), which eliminates artifacts created by orthography rather than vocabulary distribution.

The problems presented by *thorn* and *eth* and by ⁊, *and* and *ond* are just a subset of the larger challenge of handling morphological, dialectal and grammatical variants. For example, the program counts separately *cyning*, *cyninge* and *cyninga*. Additionally, variation in the spelling of diphthongs (for example, *eo* or *io*) or vowels (*i* or *y*) could influence word counts. We can use our software to lemmatize every word in a text, but this work is both time-consuming and inevitably subjective—problems we try to avoid by the use of information-processing tools. We could also *normalize*, retaining grammatical variants but consolidating spellings, but again, the process would be both time-consuming and subjective. Furthermore, we have some evidence that the distribution of various inflected forms of words can be significant, so lemmatizing them could obscure important patterns. In most of the cases we have studied, analyses performed with un-lemmatized texts yield results that are to controls (which

[23] The spelling of *and* or *ond* can be an indicator of dialect. See Campbell 1959: 110–112.

[24] For example, the scribe of *Daniel* uses *and* while the scribe of *Azarias* uses *ond*, thus creating a difference between the texts that is consistent but of only trivial interest for our purposes.

are derived from traditional philological analyses),[25] suggesting that for these particular types of problems, lemmatization is not necessary.[26] Scholars can use the software to further test the utility of lemmatization in a variety of circumstances, and it may be that lexomic analysis using lemmatized texts has the capability of providing information that is otherwise unavailable,[27] but at the present time we have found no benefit from lemmatizing.

3 Manuscripts, Editions and Editorial Practices

We must also address the problems associated with using edited or normalized texts instead of diplomatic editions. Although the Dictionary of Old English uses the most authoritative editions of Anglo-Saxon texts, these are still editions, often collated from multiple manuscripts according to the judgments of various editors, each of whose editorial practices and judgment might differ both from each other and from contemporary (and future) views. As Peter Stokes has noted, large-scale analysis using any electronic corpus can theoretically be shaped—to an unknown degree—by the editorial practices and assumptions embedded in that corpus (Stokes 2009). Furthermore, because lexomic analysis is based on critical editions, it may at times not engage particularly closely with any given manuscript. It may have been forty years since Paul Zumthor asserted the authority of manuscripts over critical editions by calling attention to the *mouvance* of medieval manuscript

[25] The exception is *Beowulf*, in which the spelling and orthographic variations between the A and B scribes are so consistent that they obscure any other potential patterns. We have dealt with this challenge by using a *normalized* text that makes spelling consistent without erasing morphological and grammatical variation through lemmatization.

[26] Scott Kleinman is currently investigating the effects of lemmatization on dendrogram geometry.

[27] For example, it may be that full lemmatization of text will allow us to apply lexomic analysis to texts on opposite size of the divide between Old and Middle English.

texts (Zumthor 1987), but even though Anglo-Saxon studies never adopted extreme points of view like Bernard Cerquiglini's assertion that "l'écriture médiévale ne produit pas de variantes, elle est variance" ("medieval writing does not produce variants, it is variance;" Cerquiglini 1999),[28] the potential significance of manuscript variation has become more important in recent years.[29]

By relying primarily on a *DOE Corpus* made up of critical editions, lexomic analysis goes somewhat against the grain of manuscript-centric approaches. It is therefore important for us to investigate the influence of both manuscript variation and editorial practice. These problems are more difficult than those of orthographic variation (which lends itself to substitutions that are easy to perform on electronic texts), but their solutions also have some fundamental similarity: by analyzing texts whose structure is already known and comparing these results with those based on manuscripts, we can see how influential both manuscript and post-manuscript variation are on dendrogram geometry.

The most significant problem is that of editorial collation. To give just one example, the *DOE Corpus* version of the *Rule of St Benedict* is based on Arnold Schröer's 1885 edition, which he produced by collating five manuscripts dating from the end of the tenth to the beginning of the twelfth centuries. Schröer's text, therefore, may not reflect any single extant manuscript or even the state of any one copy of the Old English *Rule* in any given time period (Schröer 1965 [1885]).[30] Before we put too much stock in the authority of any manuscript version of the text, however,

[28] For a useful analysis of these issues see Millett 2008, and for further discussion, see Drout & Kleinman 2010.

[29] Among the most successful applications of a manuscript-focused approach is Katherine O'Brien O'Keeffe's in *Visible Song*, in which she uses careful examination of manuscripts to demonstrate that in the Old English tradition "an oral poem did not automatically become a fixed text upon writing" (1990: 46).

[30] See also Cameron & Frank 1973: 121–122.

we should remember that the aim of Schröer's collation was to produce the most accurate possible text from a variety of witnesses, each of which was imperfect in its own way.[31] If, for example, we are interested in studying the sources of the Old English *Rule*, we would want to work with a text as close as possible to the original translation rather than a later witness in which useful information might be obscured by textual corruption. A diplomatic edition is not *a priori* more useful than a critical one.

For nearly all poetic texts in Old English the problems of editorial collations are insignificant because most Anglo-Saxon poems appear in only one copy. Furthermore, the editors of the Anglo-Saxon Poetic Records (*ASPR*) were extremely careful in their transcription and generally judicious in their emendation. Nevertheless, it may be useful to compare dendrograms produced from the *DOE*-adopted *ASPR* critical edition with those produced from a diplomatic text to attempt to gauge the significance of editorial emendation. To produce electronic diplomatic editions of our control poems, our colleague Scott Kleinman modified the *DOE*'s electronic files to make them match the manuscript forms given in the *apparatus criticus* of the *ASPR* editions. We then used these electronic diplomatic texts to repeat the experiments that had distinguished *Genesis A* from *Genesis B* and matched *Azarias* with the correct section of *Daniel*. Figure 2 shows the results of our analysis of *Genesis*. Both the diplomatic and critical editions have the same high-level clade structure in which the first major division separates *Genesis B* from *Genesis A* and the second high-level division separates *Genesis A* into two large clades, one containing segments 1, 5, 6 and 7 and the other containing segments 8 through 11.

[31] As Tom Shippey notes, it is easy to celebrate the variant *after* the production of readable editions, but quite another thing to try to puzzle out unedited texts for the first time (Shippey 2007: 151–152). See also Shippey 2008.

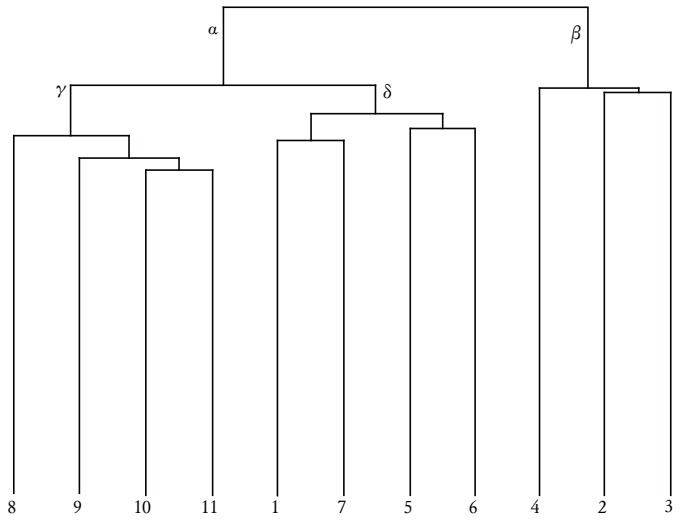

Figure 2. *Dendrogram of the* Dictionary of Old English Corpus *edition of* Genesis *cut into 1500-word segments*

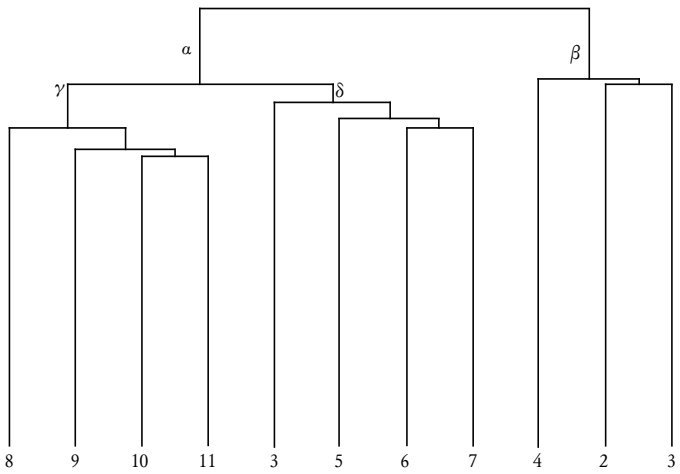

Figure 3. *Dendrogram of a diplomatic edition of* Genesis *cut into 1500-word segments*

Dendrograms of the diplomatic and critical editions are identical at the higher levels of the clade structure—those at which we separated *Genesis A* from *Genesis B*. Only deeper within clade δ do we seem some very minor variation. In both texts segments 6 and 7 are the most similar, but in the critical edition segments 5 and 1 are separately paired, while in the diplomatic edition they join the 6–7 clade in a stepwise fashion. This is actually a very subtle distinction, probably caused by small differences in segment 5 between the diplomatic and critical editions. We generally have not relied heavily on the exact geometry of the deeper clade structure, which we believe to be very sensitive to minor variations, and the results of this experiment support that approach. Since there is no difference in the high-level clade structures of the two editions, there is no reason to prefer the diplomatic edition over the critical (or vice versa) in cases where this upper-level structure is of interest. Whether we had used a diplomatic or a critical edition, we would still conclude that *Genesis A* is distinct from *Genesis B*, and indeed, these two sections have different sources.

The poems *Daniel* and *Azarias* allow us to look at a relationship of affinity. *Azarias* is quite similar to the third 900-word section of *Daniel* because both derive from the same recent antecedent Old English source even though the poems are found in two different codices, the Exeter Book and the Junius Manuscript. As we did in the *Genesis* experiment, we compared the dendrograms created using the electronic Dictionary of Old English critical editions to Scott Kleinman's reconstructed diplomatic editions.

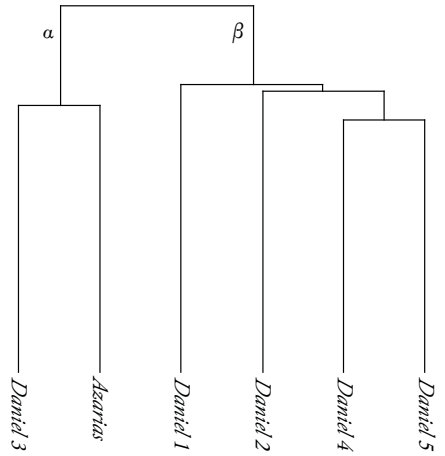

Figure 4. *Dendrogram of* Daniel *cut into 900-word segments and* Azarias *in one 1064-word segment using the* Dictionary of Old English Corpus *editions*

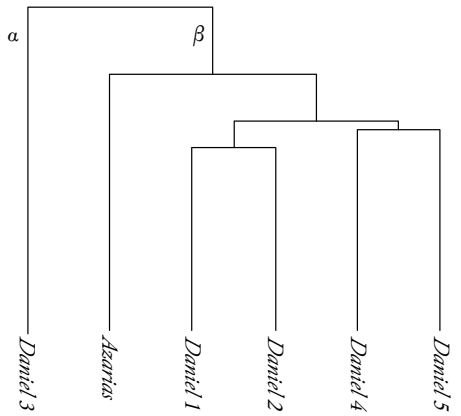

Figure 5. *Dendrogram of* Daniel *cut into 900–word segments and* Azarias *in one 1064-word segment using diplomatic editions*

In comparing Figures 4 and 5, we see that both dendrograms separate *Azarias* from *Daniel* and identify the correct 900-word segment, *Daniel* 3, as being most similar to *Azarias*. The larger

clade structure of the dendrograms is essentially the same: the first, second, fourth and fifth segments of *Daniel* are similar to each other, and *Azarias* is an outlier along with the third segment of *Daniel*. Minor differences in the two dendrograms are found deeper inside the clades. In the dendrogram created from the diplomatic edition, the *Azarias* and *Daniel* 3 leaves are separate from each other as well as from the main text, while in the dendrogram created from the critical edition, they stick together. On the other hand, within the main body of the poem segments 1 and 2 are paired in the diplomatic edition but are slightly separated in the critical edition. Based on what we know of *Daniel* and *Azarias* from the use of traditional methods—including simply comparing the poems line-by-line and word-by-word—we conclude that the dendrogram created from the critical edition is more consistent with the actual relationship between the two poems. *Azarias* is very much like the third segment of *Daniel*, and both of these are less like the rest of the poem.

Additional experiments with other texts whose sources and affinities are known (*Christ III*, the signed poems of Cynewulf, and others)[32] show that dendrograms produced from diplomatic editions of Old English poems are consistently identical at the high levels of the clade structure with those produced from critical editions. All variations that do occur are deep in the clade structure and have all been the replacement of pairings in the critical edition with stepwise arrangements in the diplomatic. Because our previous research has shown that accurate lexomic analysis is possible even when we use only those words which appear in every segment of a poem (thus eliminating the influence of rare words; Drout *et al*. 2011: 314–315), we conclude that the differences between diplomatic and critical editions are relatively invisible to lexomic methods. Because we are comparing the distribution of between 500 and 1000 words per segment, because the most common words in

[32] With the exception, as always, of *Beowulf*.

Anglo-Saxon are those least likely to be emended, and because the *ASPR* editors were extremely judicious in their textual changes, we conclude that—with the possible exceptions of *Beowulf*, *Exodus* and *Christ and Satan,* which for various reasons are heavily emended—we would gain little or nothing from replacing the electronic critical editions with re-constructed diplomatic ones. In fact, in most of the cases we have studied, the critical editions appear to be somewhat closer to the structure of the poems. We conclude that the construction of electronic diplomatic editions for the purpose of lexomic analysis is not likely to produce benefits commensurate with the effort required to produce them. However, in cases where diplomatic electronic editions do exist, it may be worth examining them as well.

It is more difficult to determine if we can have the same confidence in lexomic analyses of *prose* texts from the *DOE Corpus*. In contrast to the poems, many of the prose texts are extant in multiple manuscript witnesses. Although researchers could use the apparatus of Schröer's edition of the *Rule of St Benedict* to reconstruct all five texts as electronic diplomatic editions, the vast amount of tedious manual labor required for such an experiment is currently beyond the resources of our research group (and probably beyond the interest of all other research groups). Fortunately, Allen J. Frantzen generously provided us with electronic editions of all the manuscript versions of the Anglo-Saxon penitentials, thus enabling us to compare dendrograms derived from multiple manuscripts, both to each other and to the *DOE* edition of the text. This exercise has allowed us to determine the degree to which collation and editorial practice influences dendrogram geometry.

We chose to focus on the *Old English Penitential,* a tenth-century Anglo-Saxon text that is primarily a translation of a ninth-century Latin penitential written by Haltigar, bishop of Cambrai (Frantzen 1983: 134–139). Books 1–3 of the four-part Anglo-Saxon text translate Books 3–4 of the six-book Latin penitential (Schmitz

1958 [1898]: 275–300), but the final book of the *Old English Penitential* stems from a source written in Anglo-Saxon, the tenth-century penitential now known as the *Scriftboc*.[33]

The *Old English Penitential* is found in four manuscripts: Cambridge, Corpus Christi College, MS 190;[34] Oxford, Bodleian Library, Junius MS 121;[35] Oxford, Bodleian Library, Laud Misc. MS 482;[36] and Brussels, Bibliothèque royale, MS 8558–8563 (Catalogue number 2498).[37] These were used by Josef Raith to produce his 1933 collated critical edition (Raith 1964 [1933]), which is currently the text in the Dictionary of Old English Corpus. Frantzen's digital edition of the penitentials at http://anglo-saxon.net includes all four manuscripts. Because the amount of material in the Brussels manuscript is very small, we omit this manuscript from the following discussion.

[33] This text has, since Benjamin Thorpe's 1840 edition, been incorrectly identified as the *Confessionale Pseudo-Egberti* (Thorpe 1840). Robert Spindler also used this title for his 1934 edition (Spindler 1934), which is used in the Dictionary of Old English corpus. However, as Frantzen notes, the attribution to Egbert is found only in the incipit of CCCC 190, and the ascription most likely refers only to the "Confessional" that follows the incipit, not the *Old English Penitential* itself. In order to reduce confusion between Latin and Old English documents, Frantzen re-named the text *Scriftboc* in *The Literature of Penance in Anglo-Saxon England* (Frantzen 1983: 133–135).

[34] Ker, *Catalogue*, no. 45B; Gneuss, *Handlist*, no. 59, an Exeter manuscript from the middle of the eleventh century.

[35] Ker, *Catalogue*, no. 338, Gneuss, *Handlist*, no. 644, a Worcester manuscript from the last quarter of the eleventh century.

[36] Ker, *Catalogue*, no. 34, Gneuss, *Handlist*, no. 656, a Worcester manuscript from the middle of the eleventh century.

[37] Ker, *Catalogue*, no. 10: Glosses, penitential collections; Gneuss, *Handlist*, no. 808, a three-part manuscript containing material from the tenth, eleventh and twelfth centuries.

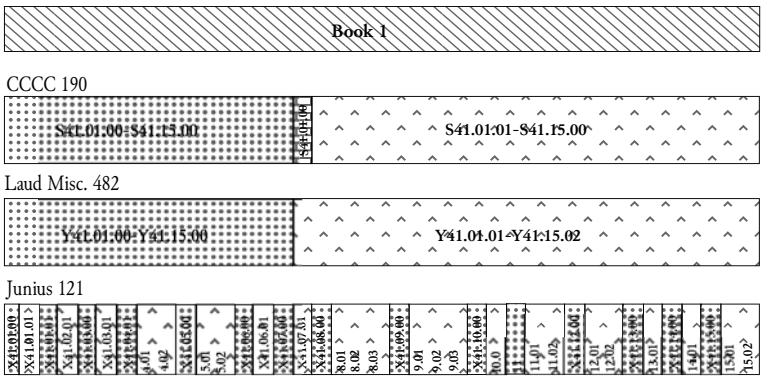

Figure 6. *Ribbon diagram of Book 1 of the* Old English Penitential *in three manuscripts*

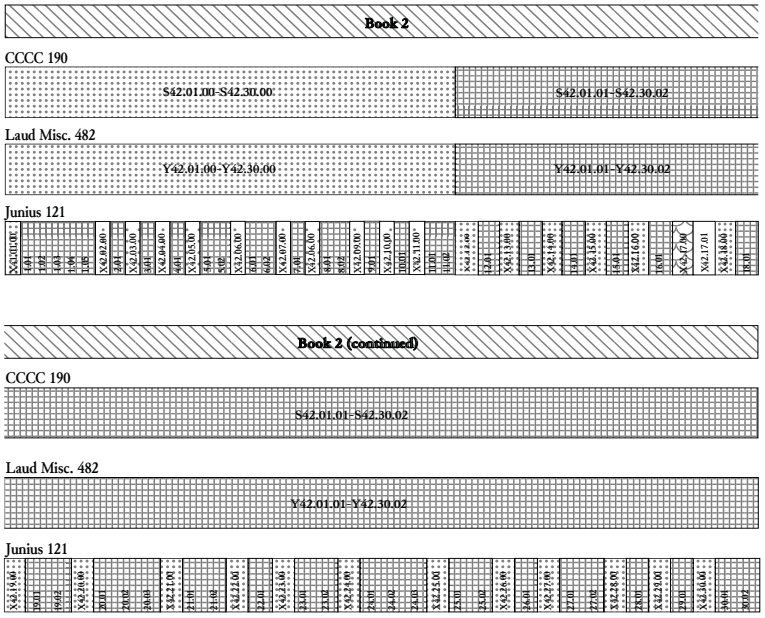

Figure 7. *Ribbon diagram of Book 2 of the* Old English Penitential *in three manuscripts*

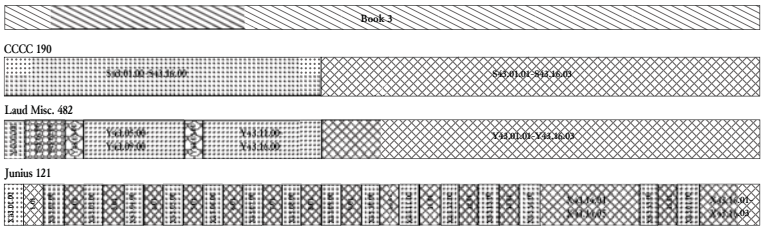

Figure 8. *Ribbon diagram of Book 3 of the* Old English Penitential *in three manuscripts*

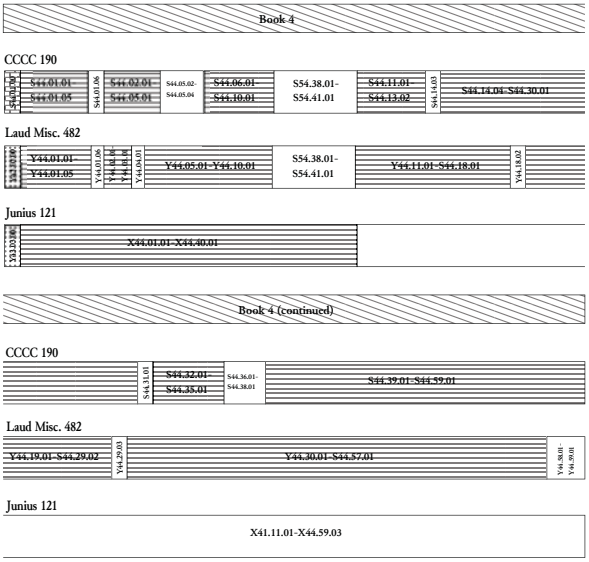

Figure 9. *Ribbon diagram of Book 4 of the* Old English Penitential *in three manuscripts*

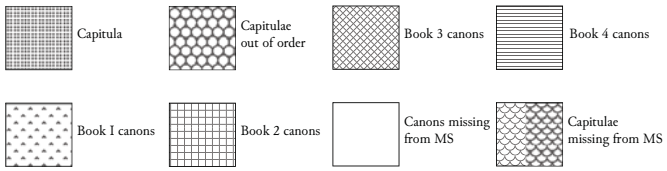

Figure 10. *Legend for ribbon diagrams*

Figures 6–9 represent the relationship among the manuscripts in what we call a *ribbon diagram*.[38] The top ribbon indicates the books (1–4) of the penitential, while the lower ribbons represent the arrangement and relative size of capitulae and canons in each of the three manuscripts. Missing and disarranged sections are indicated by shading. Notice that for books 1, 2, and 3, the ribbons for CCCC 190 and Laud Misc. 483 match up almost exactly, indicating that books 1, 2 and 3 are arranged the same in these manuscripts, with the capitulae grouped together at the beginning of each book as a table of contents. The version of the *Old English Penitential* in Junius 121, however, is differently organized, with capitulae interspersed throughout the text, as chapter headings for the canons.

Having the capitulae spread throughout the Junius text creates some challenges for lexomic analysis. Although Corpus and Laud match up segment by segment regardless of segment size, the same content is distributed somewhat differently in the Junius manuscript: the first 1000 words of Laud and Corpus are made up entirely of capitulae, while the first 1000 words of Junius are approximately 65 percent text and 35 percent capitulae. To address this problem we used a process we call *blending*[39] to re-arrange the material in the CCCC 190 and Laud manuscripts in order create segments that would allow one-to-one comparisons. We therefore cut the first three books of Corpus and Laud between the capitulae and the main text and then sub-divided each of these segments in half, producing for each book two shorter segments composed entirely of capitulae and two short segments composed entirely of regular text. We then matched the first segment of capitulae with the first segment of text, the second segment of capitulae with the second segment of text, and so on, and then blended together the capitulae and their now-associated main text into new segments. Figure 11 illustrates the process.

[38] Ribbon diagrams were developed by M. D. C. Drout and Courtney LaBrie in 2011.

[39] The blending technique was developed by M. D. C. Drout and Leah Smith.

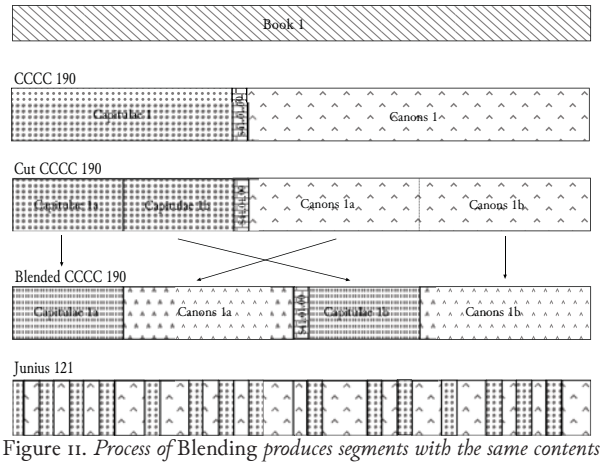

Figure 11. *Process of* Blending *produces segments with the same contents*

Because the capitulae run in order, putting the first half of the CCCC 190 and Laud capitula sections with the first half of the corresponding text creates new, hybrid segments that are composed of the same material as the Junius text segments, in which the capitulae are interspersed. We can then use these texts to create dendrograms of the three manuscript witnesses of the *Old English Penitential*.

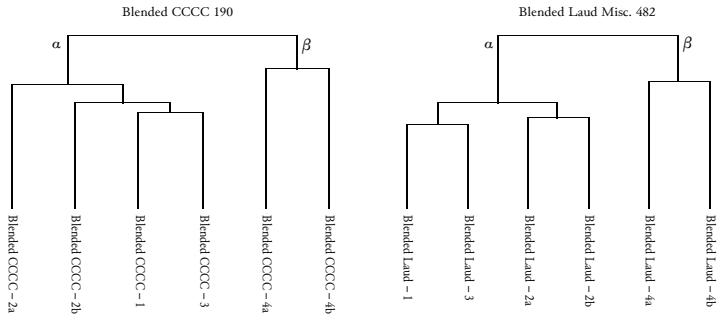

Figure 12. *Comparison of dendrograms of the* Old English Penitential *in CCCC 190 and Laud Misc. 482, segments blended*

Of the three manuscripts, Corpus and Laud have the most similar dendrogram geometries, and in the highest level of the clade structure they are the identical. In Figure 12 the segments are named by their relationship to Books of the *Old English Penitential*. Books 1 and 2 are complete in individual segments; Books 2 and 4, because they are larger, are each divided into two segments, "a" and "b." The high-level clade structure of the texts in the two manuscripts is identical: segments 1, 2a, 2b and 3 cluster in one clade and segments 4a and 4b in the other. Furthermore, this high-level clade structure is consistent with what we know of the sources of the *Old English Penitential*: clade α (segments 1, 2a, 2b, and 3), on the left of the dendrogram, has Haltigar's Latin penitential as its source; the material represented by clade β (segments 4a and 4b), on the right side of the dendrogram, is taken from the Old English *Scriftboc*.

In both manuscripts, segment 1 clusters with segment 3, but in Laud, segments 2a and 2b also cluster together, while in Corpus 190 we see a stepwise geometry with 2a and 2b slightly separate. Because the vertical distances between the branches are so short between the inner clades, the geometry may be perturbed by only very small variations in the underlying text.

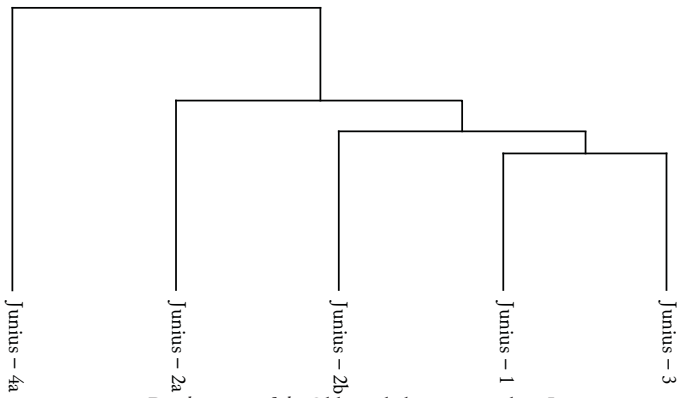

Figure 13. *Dendrogram of the* Old English Penitential *in Junius 121*

Although the Junius dendrogram in Figure 7 at first glance appears to have a geometry different from that of the Corpus and Laud dendrograms, closer examination shows that the dendrograms are the same as long as we take into account the absence of some material from the Junius text. In all three dendrograms, segments 1 and 3 cluster most closely, then segments 2b and 2a join that clade (stepwise in Junius and CCCC, pairwise in Laud). Material from the fourth book differs most in vocabulary and is thus separate from the rest of the dendrogram. This clade is simplicifolious in Junius simply because the text corresponding to segment 4b in Laud and CCCC is missing from the manuscript. And, as Figure 8 shows, the Junius dendrogram also has essentially the same geometry as the *Dictionary of Old English* collated text, with the only difference being the absence of segment 4b. This geometry is explained by Raith's editorial practice of using material from Junius to fill in gaps in Corpus and Laud. Raith's combined text therefore makes segment 4a somewhat different from what it is in either Laud or Corpus (where 4a is more similar to 4b).

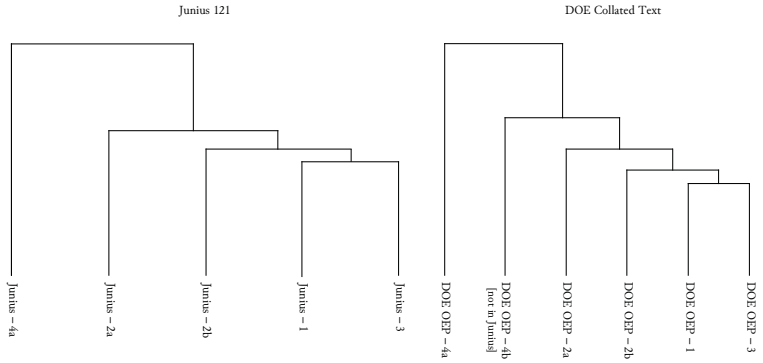

Figure 14. *Comparison of dendrograms of the* Old English Penitential *in Junius 121 with the* Dictionary of Old English *collated edition of the same text*

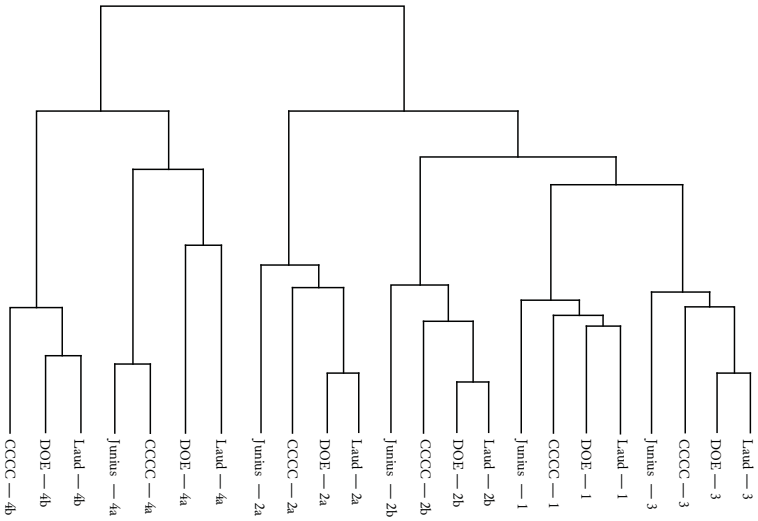

Figure 15. Comparison of dendrograms of the Old English Penitential *in the Laud, CCCC, and Junius 121 manuscripts and the* Dictionary of Old English *collated edition*

Figure 15 compares the *Dictionary of Old English* critical edition with all three diplomatic editions. As we would by now expect, lexomic methods correctly place matching segments together (even though the texts are not entirely identical). We also see that within clades, the segments of the *DOE* collated critical edition stick most closely to the corresponding segments of the Laud manuscript, showing that the *DOE* edition follows the vocabulary of the Laud manuscript more closely than it does the other manuscripts. If we simplify the terminal leaves of the dendrogram (Figure 16), we can more easily see how the relationships between the texts and the critical edition are represented in the higher-level clade structure.

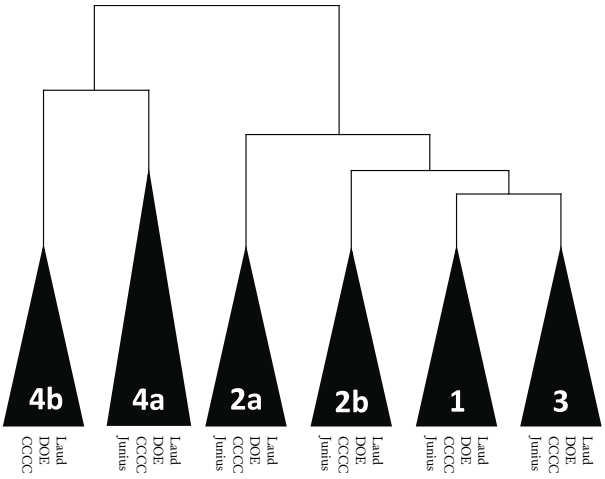

Figure 16. *Comparison of dendrograms of the* Old English Penitential *in the Laud, CCCC, and Junius 121 manuscripts and the* Dictionary of Old English *collated edition, terminal leaves simplified*

It is now easy to see that the combined dendrogram has the same high-level clade structure as the Junius dendrogram in Figure 13 (again with the exception that segment 4b is absent from the Junius text).

From these experiments we can draw several conclusions. First, at the higher levels of the clade structure, there is no significant disagreement between the dendrograms produced from diplomatic and critical editions. We can therefore use either and still get results that agree with the controls. Furthermore, we note that the relationships of source structure that are of particular interest to us are represented in the dendrograms of the prose texts regardless of manuscript or edition. In all cases, the material with an Old English source is separated from that with a Latin source at the highest level of the clade structure. There are small differences in dendrogram geometry between diplomatic and critical editions at lower levels of the clade structure, but these are subtle, in each case being the difference between a stepwise and a pairwise arrangement of clades

with very short vertical distances between them, a geometry that indicates only small differences in vocabulary that should not be used to draw significant conclusions. If the ultimate exemplar of the *Old English Penitential* included the first three books translated from Haltigar plus a fourth book taken from the Old English *Scriftboc* (the conclusion arrived at using traditional methods), then the critical edition accurately reflects this archetype. Furthermore, the dendrograms made from the critical edition display the same basic clade structure as those from the diplomatic editions of the manuscripts.

Our reception of texts from before the age of mechanical reproduction is strongly influenced by editorial practices, many of which are opaque to us if we read a text for content alone. We must therefore pay close attention to editorial practices at every level, from orthography to word division, emendation and collation, all of which have the potential to affect the data we are using to produce dendrograms (and thus analyze textual structures and relationships). Prose texts, which are longer and often exist in more manuscript witnesses than Old English poetic texts, present challenging problems, especially if we want to compare those edited by different editors, whose practices likely vary. However, based on both our previous analysis of Anglo-Saxon poetry and the results of this examination of the editions of the *Old English Penitential*, we can have reasonable confidence in lexomic investigations that use the critical editions in the *Dictionary of Old English Corpus*.

4 Lexomic Detection of Sources: The *Old English Orosius*

The Spanish priest Paulus Orosius wrote his *Historiarum adversus paganos libri septem* in 417 or 418 at the urging of his teacher, St Augustine of Hippo. This universal history, which covers events from the Fall of Man to the early fifth century, was polemical as well as historical, attempting to demonstrate that the political and social disruptions of the author's lifetime were not due to the adoption of Christianity and subsequent neglect of the pagan gods.

In the past there were even more disasters, Orosius asserts, and so Christianity need not shoulder the blame for current (fifth-century) conditions. The *Historiarum adversus paganos* was extremely popular throughout the Middle Ages, with over 250 surviving manuscript witnesses (Bately 1980: lv). Some time between 889 and 899 and probably closer to 890, Orosius's Latin text (hereafter abbreviated as *OH*) was translated into Anglo-Saxon (Bately 1980: lxxxvi–xciii). Based on the testimony of William of Malmesbury, this Old English translation (abbreviated *Or*) was traditionally attributed to King Alfred (Stubbs 1887: I.132). Alfred's authorship was never seriously questioned until 1951 (Raith 1951: 54–61), and it was only in 1970 that Janet Bately demonstrated that the translation was almost certainly not by the king himself, although it is likely to have been produced as part of Alfred's educational and translation programs (Bately 1970: 433–460).

Where it follows the source text the Anglo-Saxon translation is a basically accurate rendering of *OH*, but as Bately notes, the translator does not hesitate to omit or reduce the description of many of Orosius's interpretations of events, at times replacing them with his own observations or analyses and on the whole converting the text from a polemical document addressing a fifth-century audience to a more general "survey of world history from a Christian standpoint" (Bately 1980: xciii). The translator also augmented his text with incidental material from various classical and patristic authors[40]—perhaps drawn from annotations in the Latin manuscript that was his exemplar or from commentaries—and with geographic information not present in *OH*. The most famous of the additions are the reports of ninth-century voyagers Ohthere and Wulfstan (hereafter be referred to as the Voyages), which describe the lands and cultures of the north, but there is also

[40] The most complete and up-to-date list of identified or suspected sources can be found in the *Fontes Anglo-Saxonici: World Wide Web Register*, http://fontes.english.ox.ac.uk (accessed 25 February 2013). See also Bately 1971 (but note that this important article is keyed to Sweet's edition, not Bately's later text).

a great deal of geographic material in *Or* which either replaces or augments the contents of *OH*.[41]

The *DOE Corpus* electronic text is Bately's definitive 1980 E.E.T.S. edition, which is based upon London, British Library, Additional MS 47967 (manuscript L), except for section 15/1–28/11, which are missing in L but found in London, British Library, Cotton Tiberius MS B.i. (manuscript C). Although Bately adopts a few additions and corrections from other manuscripts (indicated by square brackets in her text), her edition is not an artificial conglomeration of multiple sources but a judicious reconstruction of the single manuscript that seems closest to the original Old English archetype (Bately 1980: xxxviii–xxxix concludes that MSS L and C are at least two removes from that text). Our pre-processing for lexomic analysis, then, only requires that we consolidate *thorn* and *eth* and lemmatize Tironian, *and* and *ond* as well as performing our standard "scrubbing" to remove formatting and punctuation and force all letters to lower-case.

We chose to divide the text into 900-word segments, a size which requires some explanation. Previous research has shown that dendrograms of Anglo-Saxon poems are broadly accurate down to a segment size of 500 words, but that dendrograms based on segments closer to 1000 words somewhat more consistent in detail with the known structures of texts (Drout *et al*. 2011: 311–315). The trick is to avoid creating segments that split apart significant features of a text (for example, spreading the *Azarias* section of *Daniel* across two segments) and therefore producing artifacts in the dendrogram. Unfortunately, when we are dealing with a text whose sources are unknown or only suspected we do not have a source-structure to guide the arrangement of our dividing lines. In these cases we have found it useful to create multiple dendrograms of varying sizes to see

[41] The possible sources of the geographic material has been the subject of an enormous amount of scholarship. See Bately 1980: lxiii–lxx, lxxxix–xc. The most recent discussion, which is extremely thorough, is Valtonen 2008.

what patterns are *robust* across different segment sizes (for example, we might start with 800-word segments and then increase the segment size by 100 words until we reach 1500-word segments). We are then able to isolate distinctive sections of a text by making small adjustments in segment boundaries in subsequent experiments.[42] Because the Voyages of Ohthere and Wulfstan are a known feature of *Or*, we sought to avoid combining these with other material in a single segment in order to avoid creating a hybrid whose vocabulary distribution was representative of neither the Voyages nor the non-Voyages material. A 900-word division puts the Voyages into two segments (3 and 4) that do not include non-voyage material. The dendrogram that results from performing cluster analysis on the 900-word segments of the scrubbed Old English text is shown in Figure 17. There are fifty-five 900-word segments.

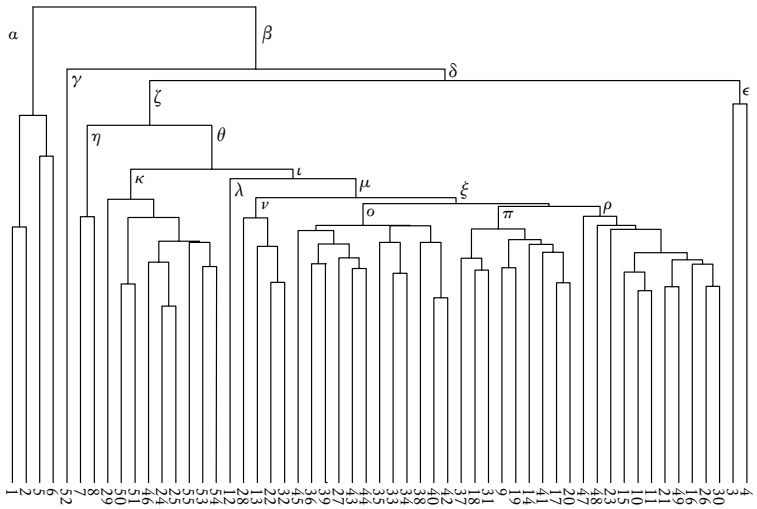

Figure 17. *Dendrogram of the* Old English Orosius *cut into 900-word segments*

[42] Because we calculate all distances using relative frequencies, it is not essential for the absolute sizes of each segment to be identical. However, we have found it important to avoid extreme differences in segment size because very large differences have the potential to produce artifacts in the dendrograms.

When faced with as large and complex a dendrogram as Figure 17 (a situation more likely in the analysis of prose texts than of shorter poems), it is useful to bundle together the terminal leaves of many clades in order to see more clearly the high-level clade structure. Like Figure 16 above, Figure 18 borrows a convention from linguistics and represents large clades with triangles. The high-level clade structure of the dendrogram is thus seen to be relatively simple. There is a very significant divide between clade α (which contains only segments 1, 2, 5 and 6) and the much-larger β, which includes all the rest of the Orosius translation. There are then four major divisions in β: single-leafed γ, and the bifolious clades ϵ and η are distinct from θ, which contains 46 segments.

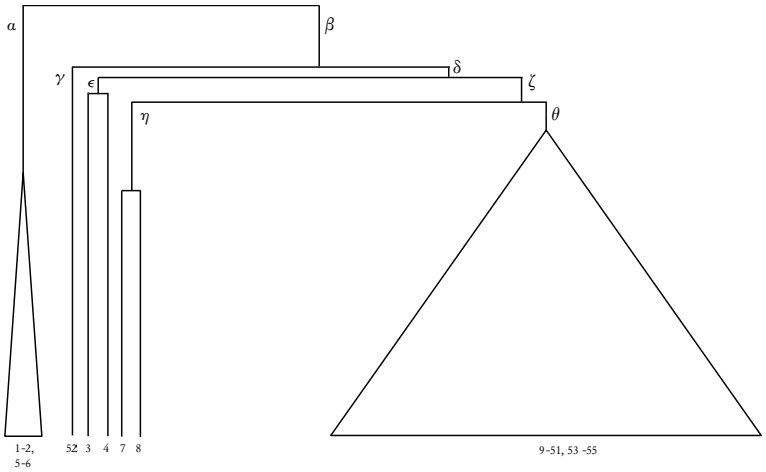

Figure 18. *Simplified dendrogram of the* Old English Orosius *cut into 900-word segments*

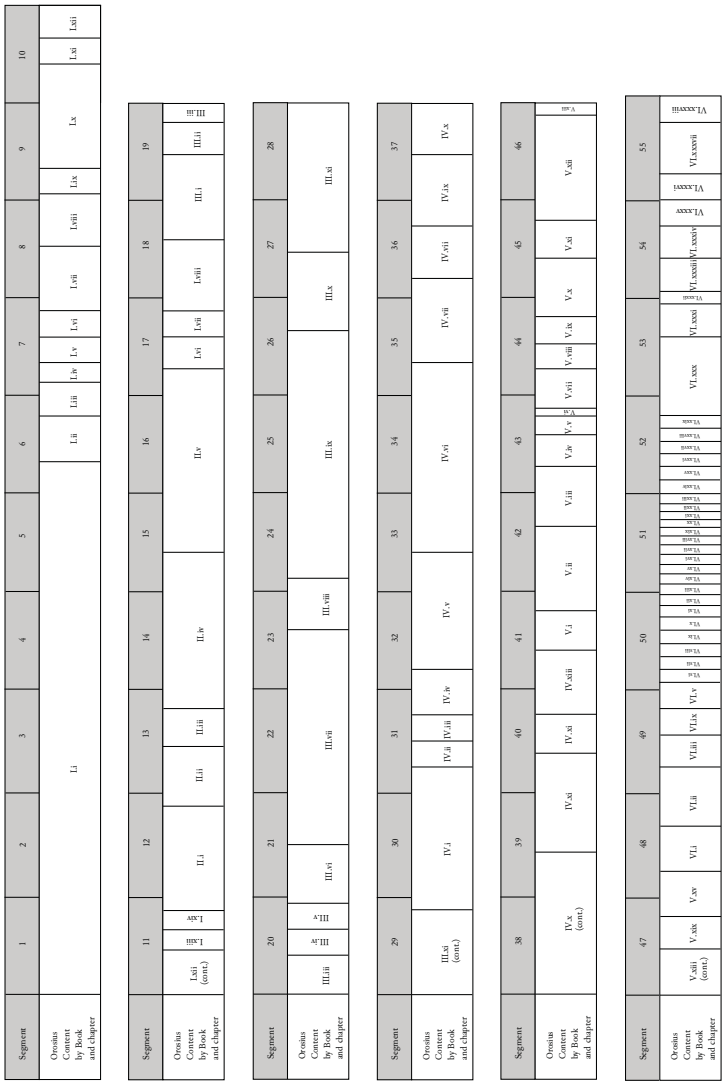

Figure 19. *Ribbon diagram of the* Old English Orosius *shows the organization of the text and the relationship of that organization to the segments of the dendrogram*

The ribbon diagram in Figure 19 can be used to correlate the placement of each segment in the dendrogram with its content.[43] The top row gives the segment number, the bottom row the book and chapter in the *Old English Orosius*. The core of the dendrogram, clade θ in Figures 17 and 18, represents a significant quantity of material—over 41,000 words—translated from the Latin text of Orosius's History. The short vertical distances between sub-clades indicates that the material in this large grouping is relatively homogenous (though there are some differences, to which we will return). As is discussed in much more detail below, clades α, ϵ and η, which are separate from θ, all have different sources than the main body of the text. Most significant for our present purposes is clade ϵ, which contains the Voyages (pages 13–18 in Bately's edition). Originally written in a vernacular and thus not translated from Latin, the Voyages have long been noted to be linguistically different from the rest of the *Old English Orosius* (Bately 1980: lxxii). They also differ from each other. Ohthere's account is that of a Scandinavian visiting England, whereas Wulfstan's is of an Anglo-Saxon who had traveled to Scandinavia (Townend 2002: 90–95). Although the degree of influence of Old Norse upon Ohthere's account is disputed (Townend 2002: 95–101), there is no doubt among scholars that the Voyages were composed in Old English. It is therefore significant that they are so distinctly separated from the main body of the dendrogram. As in our lexomic analysis of the *Old English Penitential*, we are able to detect sections of a text that have significantly different sources than those of the main body of the text. Even at this relatively crude level of analysis, therefore, we have taken a significant step towards establishing the accuracy of lexomic methods for Old English prose, since the placement of segments 3 and 4 indicates that these have a different source from the other segments, which they do.

[43] For a much more detailed breakdown of the contents of each segment, see Appendix A.

There are, however, additional separated clades in the dendrogram that do not contain material from the Voyages and therefore require further analysis. Clade α contains segments 1, 2, 5 and 6 of *Or* (Book I, chapters i–iii, with the exception of the Voyages). The sources of this geographic material are unknown and disputed. Some of the geographic information may have been drawn from a *mappa mundi* (Derolez 1971),[44] and other elements appear to come from the translator's general knowledge of continental Europe in the ninth century (Bately 1980: lxvii–lxx). But regardless of where the material came from, it is certain that it is not drawn from the Latin text of Orosius's history. Thus clade α, like clade ϵ, also has a different source than the main body of the text in clade θ, and this difference is reflected in its placement in the dendrogram.

Clade η likewise has additional sources beyond *OH*. This bifolious clade is comprised of segments 7 and 8, which contain the last third of chapter iii and chapters iv–viii of Book I. As Bately demonstrates in her commentary, the material in this section is heavily modified and augmented from the Orosius's original. For instance, Bately notes that at the end of chapter iii, a comment derived from Josephus (by way of Hegesippus) has been interpolated into the text. Although the comment is found in various manuscripts of *OH*, it is absent from those that are closest to the deduced source of the *Old English Orosius*. Its inclusion, therefore, suggests that the translator used an additional source here, perhaps Isidore's *Etymologies* (Bately 1980: 212–213).[45] According to Bately's commentary, segments 7 and 8 contain 16–18 places where *Or* contains additional material

[44] For caveats see Bately 1980: lxvii–lxx, who does not rule out the use of one or more more *mappae mundi* but notes that the evidence of the text is "sadly inconclusive."

[45] Bately notes that the comment could have been derived from Augustine, Tertullian or Tacitus, and that the version closest in wording to the Old English text is Bede's *De locis sanctis*. The *Fontes* database identifies Bede as the most likely source.

not found in Orosius's Latin text.[46] In comparison, Bately only identifies five unambiguous and two possible additions in segment 9, and these are shorter than those in segments 6 and 7.

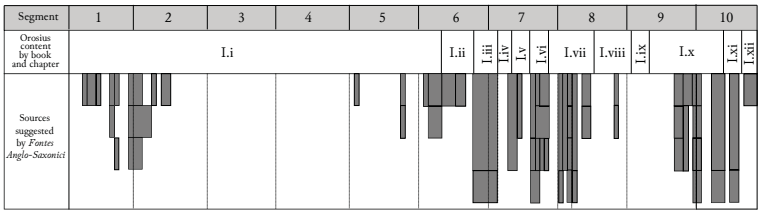

Figure 20: *Ribbon diagram of the identified sources (based on the* Fontes Anglo-Saxonici *database) of segments 1–10 of the Old English translation of Orosius's* Historia. *Segments are 900 words long. Note the lack of Latin sources for segments 3 and 4*

The *Fontes Anglo-Saxonici* database adds somewhat to this total: of the 165 lines in segments 6 and 7, *Fontes* identifies 72 of them as having sources in addition to Orosius, and approximately 20 of these lines are definitely *not* from that source. Of the 87 lines of segment 9, the *Fontes* database identifies up to 35 as possibly having a source outside of Orosius, but none of these definitely has an outside source. Figure 20, which represents the information in the Fontes database, shows that indeed there are more potential sources in segments 7 and 8 than in segments 9 and higher.

But close inspection of the citations in the *Fontes* database suggests that we must be somewhat cautious here: the database lists all the possible sources for a given line but often does not indicate which is the proximate source for the Old English translator because many of the citations are to the use of ideas rather than to phrasing from any specific text. Although the translator might have been consulting a florilegium, a well-glossed commentary or manuscripts in a well-stocked library, he may also have drawn on his own general knowledge and prior reading. For example, Bately

[46] There are 16 notes in which Bately 1980: 212–218 identifies definite additions and two other places in which she suspects an addition.

and the *Fontes* database identify *Genesis* 41:29 as the source of a substantial passage in Book I, chapter v that describes Joseph's prediction of the seven fat years (Bately's lines 23.19–24.15), a passage not found in *OH* (Bately 1980: 213). Certainly the ultimate source for the passage is the Bible, but it seems likely that here the translator is merely drawing on his memory of the story than any specific intermediate source, since the Old English does not translate the biblical text word for word. This passage therefore does have a *different* source from that of the nearby material that translates Orosius's Latin, but we cannot be certain which text was its proximate source. Many of the other identifications of sources are likewise difficult to link to a physical text. But while segments 7 and 8 appear to have more definite sources than most of the other, later segments in clade θ, the density of material from non-Orosian sources is not nearly as pronounced in this clade as it is in ϵ (the Voyages) or α (the geographic material). And indeed, although η does separate from θ, it is the closest of all the outliers to that very large grouping of segments and so most similar to the main body of the text that is translated for the most part directly from Orosius's Latin.

The remaining anomaly in the dendrogram is segment 52, which in vocabulary distribution is less distant from the main body of the text than the geographic material, but, surprisingly, more so than the Voyages. This segment contains Book VI, chapters xxiiii–xxix and half of chapter xxx. Differences between this segment and those before and after it are not readily obvious. At this point in Book VI there are a series of short chapters that have the effect of repeating the opening words "Æfter þæm þe Romeburg getimbred wæs __ wintrum" [after the time in which Rome had been established for __ years] more frequently here than in many other segments, but not particularly more so than in 51 and 53. The ribbon diagram in Figure 21 shows that segment 52 has very few identified sources (*Fontes* and Bately propose oblique influence by Jerome and Isidore, but this identification

is tentative and there are no obvious quotations). Bately's note on the end of chapter xxviii speculates that there was corruption in the underlying manuscript at this point in the text, and it seems plausible that a damaged or defective text could influence a dendrogram, but in this particular case only a single sentence appears to have been affected directly by the corruption.[47] We are therefore left without a good explanation for the placement of segment 52. Either it has a source or author different from the main body of the text but not identified by Bately or *Fontes*, or its very lack of additional external sources makes it distinctive in vocabulary (although segment 22 similarly has few or no known sources beyond that of the Latin Orosius).

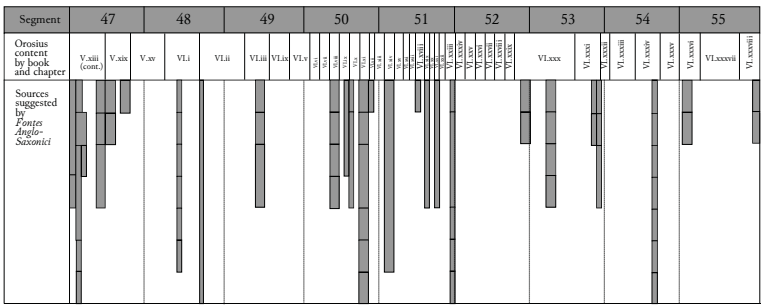

Figure 21: *Ribbon diagram of the identified sources (based on the* Fontes Anglo-Saxonici *database) of segments 47–55 of the Old English translation of Orosius's* Historia. *Segments are 900 words long. Note that 52 is the only segment generally lacking in known sources*

Segment 52 therefore is at this point an unexplained anomaly. As such, it could cast some doubt on the applicability of lexomic

[47] It may be that corruption in the exemplar affected the prose styly by forcing the translator to *compose* rather than translate and that the resultant difference in the distribution of vocabulary is influencing the dendrogram geometry, but at this stage of our knowledge we do not have enough evidence or understanding to confirm or rule out this possibility.

methods to prose texts,[48] but in all other cases the high-level of the geometry has reflected the source structure of the texts. lexomic methods were able to detect the differences in vocabulary distribution between the section of the *Old English Penitential* that has an Anglo-Saxon source and the rest of the text, which is based on the Latin penitential of Haltigar, and they were likewise able to identify through dendrogram geometry alone the influence of different sources in the Voyages of Ohthere and Wulfstan, the geographic material from an unknown source, and the additions to segments 7 and 8 of the Orosius translation. We can therefore have some reasonable confidence in the accuracy of the methods when extended from poetry to prose, particularly when we remember that we are using lexomics to open up a complimentary information channel about texts, not to replace traditional methods. It is when we correlate traditional methods with lexomic analysis, using each to augment the other, that we gain new insight into the texts, and it is hoped that future scholars, now alerted that something may be unusual in segment 52, may be able to discover an explanation.

5 The Deeper Structure of the Orosius Translation

To this point our paper has mostly developed controls to which future lexomic research into prose text can be compared. It is hard to overstate the importance of such controls in a historical discipline like ours: we can only have confidence in the techniques if we can compare the results arrived at by their employment with knowledge acquired by other means. But while controls can show us that a methodology can produce accurate results, they do not necessarily demonstrate that the methods are useful. For this latter point we want not merely confirmation of existing knowledge but

[48] Although other seeming anomalies, such as the placement of *Juliana* in the Cynewulf dendrogram or a seemingly anomalous simplicifolious clade in *Genesis* have turned out, upon further study, to have external sources. See Drout *et al.* 2011: 330-335.

unexpected additional support for more controversial hypotheses or entirely new information. Further analysis of the lower-level clade structure of the Orosius translation gives us examples of both desiderata.

For the purposes of our preceding analysis we had simplified the large and complex dendrogram of the entire translation, temporarily ignoring the details of the structure of 46-leafed clade θ (in Figure 17). It is now time to examine its geometry more closely. Within θ the first clade to separate is κ, composed of segments 24, 25, 29, 46, 50, 51, 53, 54 and 55, and within κ segment 29 is simplicifolious, indicating that the distribution of its vocabulary is distinct from the rest of the material.

Clade 29 contains the second half of Book III, chapter xi, in which Orosius discusses the struggles among the successors of Alexander the Great in Macedonia. The twists and turns of the plot are complex, with multiple treasons and shifts of fortune. As Bately demonstrates, the Old English translator attempted to make this section of the text clearer both by simplifying and by adding explanations. Bately's notes refer frequently to the *Epitome* of Justinus,[49] an early Roman historian who was a source for Orosius and whose writings clarify—at least for the modern reader—the Alexandrine succession. There is never a close enough verbal correspondence between Justinus and *Or* for Bately to be certain that the *Epitome* was a source for the translator, but reading Justinus side-by-side with both *OH* and *Or* does show how opaque some of Orosius's passages are in comparison to both the *Epitome* and the Old English text.[50] Additional circumstantial evidence for the influence of Justinus may be the possibility that Asser, King Alfred's biographer, knew the *Epitome*, since Michael Lapidge has shown that Justinus cannot be ruled out as a text known to Asser

[49] For the knowledge of Justinus in Anglo-Saxon England, see Crick 1987.

[50] We are grateful for Joel Relihan's assistance with this material.

(Lapidge 2003: 27).[51] Although the Orosius translation is no longer credited directly to the king, it is understood to have been part of his educational program and produced in his circle, of which Asser was an important part, arriving at approximately the same time as Grimbald and John the Old Saxon (Keynes & Lapidge 1983: 26–27). If Asser had access to a copy of Justinus—either in Wales or England—then it is not unreasonable to suppose that the translator of *Or* could likewise have read Justinus and therefore use the *Epitome* as a source for this section of his translation. The influence of Justinus would then explain the placement of segment 29 in the dendrogram.

However, when the Old English translator deviates from Orosius's text, he does not obviously translate Justinus. Instead, it almost appears as if the translator becomes frustrated with Orosius's circumlocutions and rather brutally simplifies the material. For instance, Orosius, in his depiction of the death of Lysimachus, offers an elaborate, somewhat poetic description, which the Old English translator renders tersely as "þær wæs Lysimachus ofslagen" (Sweet 1883: 152–153; Bately 1980: 82; Seel 1972: 148). Other, similar simplifications are found throughout the section, suggesting that the influence of Justinus, if it exists at all, is somewhat oblique.

The final passage of Book III, also included in segment 29, comes from neither Orosius or Justinus: "þonne us fremde & ellþeodge an becumaþ & lytles hwæt on us bereafiað & us eft hrædlice forlætað." As Bately notes "There is nothing corresponding to this in *OH*: indeed the situation in Rome in Orosius' day was very different. It therefore seems reasonable to suppose that the translator is referring here to conditions in his own time, and to raids by the Vikings" (Bately 1980: 270 and xciv). Here we see the translator modifying and augmenting his source based on what is presumably his own experience rather than any external text. The additional

[51] However, Lapidge was not able to confirm Asser's knowledge of Justinus because the evidence is ambiguous.

material here is only 18 words long, so it itself is almost certainly not the entire cause of the location of segment 29 in the dendrogram. But the obvious departure from the source may indicate that in this section of the text the translator was more freely adapting than elsewhere, either because he was frustrated by and wanted to clarify Orosius, or because he had a text—Justinus—that better explained the material, or some combination of the two.

Given the current state of lexomic techniques and our knowledge of the text, we cannot conclude at this time that segment 29 certainly has a different source (or constellation of sources) than the rest of the Orosius translation. But the correlation of the lexomic evidence with information derived from traditional methods of investigation gave us a reason to reexamine the evidence for changes in influence at this point in the text, and our subsequent scrutiny of the text has at least hinted at the translator's practice (and perhaps his sources and identity). Further examination of the Orosius translation in light of the geometry of dendrograms, especially those composed of differently sized segments, may reward investigators who can correlate dendrogram geometry with previous hypotheses about structure, authorship or affinity. For example, Elizabeth Liggins's 1970 claim for multiple authorship of the translation—based on her analysis of the distribution of various syntactic features—was reasonably criticized by Bately for, among other reasons, lacking a "control" and for failing to take into account "the possibility of a single translator, gradually developing a style" (Bately 1980: lxxiv–lxxxi). Lexomic analysis does not on the first pass appear to support Liggins's assertions, but it may be worth noting that the interior structure of clade θ, which contains the most homogeneous section of the translation, does divide roughly into three large clusters (clades o, π and ρ in Figure 17). In research on other texts,[52] we have found that the production of multiple dendrograms at different segment sizes allows us to note "robust" groupings (those which appear at multiple resolutions).

[52] Drout *et al*. forthcoming.

Identification of such robust divisions and then further syntactic, semantic or stylistic analysis may allow researchers to revisit the claims for multiple authorship or to gain a better understanding of the translator's practice. Although such analysis is beyond the scope of the current paper, which has been concerned to establish a baseline of knowledge about the applicability of lexomic methods to Anglo-Saxon prose, it can be performed with comparative ease now that the software tools are freely available and now can be operated through a convenient interface.

6 Conclusions

This paper set out to determine if the lexomic techniques which have been profitably applied to Anglo-Saxon poetic texts might also be used for analysis of Anglo-Saxon prose. We conclude that with suitable modification they can. Researchers must take into account not only the larger size of most prose texts but also their existence in multiple copies and recensions, which are reflected in the complexity of the critical apparatus of most editions. Our investigation of the *Old English Penitential* shows that lexomic analysis based upon a critical edition is consistent with that based on a diplomatic edition, but we also note it is essential that researchers understand thoroughly an editor's practices of collation and organization. Had we not recognized that Raith interpolated the capitulae from Junius 121 into a text based primarily on Laud Misc. 482, we would not have been able to devise a useful experiment and then interpret the dendrogram correctly. Combined with our comparison of the *ASPR* critical editions of poems to the diplomatic editions reconstructed from the apparatus, the evidence of the *Old English Penitential* dendrograms gives us some confidence in lexomic analysis based on the critical editions in the *Dictionary of Old English* corpus. It is important to note, however, that differing editorial practices across multiple texts may complicate the task of comparing them, and while the consistent editing of Krapp and Dobbie across the entire poetic corpus allows us to make comparisons among Anglo-Saxon poems, there is no such

consistency in the editions of much of the prose. If, for example, we wanted to perform lexomic analysis across all the penitential texts in the *DOE Corpus*, we might find that consistent differences between Raith's editorial practices and those of Finsterwalder or Mone might generate either false positives or negatives. Consolidation of thorn and eth and expansion of Tironian note will obviate some artifactual differences, and others can be eliminated through orthographic normalization or even lemmatization. In the end, researchers can have confidence in lexomic analysis based on any single critical edition but must be cautious when making broader comparisons.

Our analysis of both the *Old English Penitential* and the Old English translation of Orosius allows us to conclude that the ability of lexomic methods to detect significant differences in the sources of texts applies to prose as well as to poetry. The dendrograms of both the penitential and *Or* separates material based on its sources: the final book of the penitential, which is based on the Anglo-Saxon *Scriftboc,* is in its own clade, as are both the non-Orosian geographic material and the Voyages of Ohthere and Wulfstan. Furthermore, like material appears to be grouped with like: despite the interruption of the Voyages in segments 3 and 4, the most outlying clade in the Orosius dendrogram contains all segments of geographic material derived from an unknown source (segments 1, 2, 5 and 6). The dendrogram does not only separate differently sourced segments but groups them correctly.

In addition to establishing controls, this paper has set out to demonstrate the utility of lexomic analysis in Anglo-Saxon prose texts. Our discussion of the possible influence of Justinus on *Or* shows both the promise of the methods and their challenges. In this particular case, we had no particular agenda with regard to the possible use of Justinus by the translator because we were unaware that this was an open question in the scholarship. The dendrogram is therefore reasonably objective evidence that segment 29 is subtly different in vocabulary distribution from the material that surrounds it. In itself that difference is not sufficient evidence to be

certain that the translator knew Justinus. Even when we combine the lexomic evidence with Bately's very tentative hypothesis and Lapidge's identification of the *Epitome* as a text that might have been known to Asser, we still find ourselves in speculative territory. But although the accumulation of circumstantial evidence is never dispositive, it is still valuable, and we can therefore conclude that the translator's use of Justinus is somewhat more probable than it was before we knew of the lexomic results.

Perhaps more significantly, we see here that the lexomic approach can show us *where to look* even if it cannot always tell us what we end up finding there. Most investigations in our field are thesis-driven: we have a hypothesis and seek evidence to support it. Lexomic analysis can certainly be used this way, but it is perhaps even more valuable when we realize that because they are broadly objective and able to be automated, lexomic methods can be used as screening mechanisms. The Orosius translation is enormous and Bately's edition larger still. Most researchers must approach such large texts with a pre-existing thesis for which they seek supporting evidence. In such circumstances, the mind's ability to detect large-scale, unanticipated patterns is limited. Lexomic methods, however, can screen multiple large texts to identify particular sections that might repay scrutiny. Once these segments of interest are identified, scholars can employ traditional methods and then, in an "iterate and test" loop, return to lexomic approaches in order to generate additional evidence with which to test various hypotheses. Although they will never replace the erudite and creative scholar, lexomic methods do have the potential to become a significant tool for better understanding the culture of the Middle Ages.

Phoebe Boyd, Michael D. C. Drout, Namiko Hitotsubashi,
Michael J. Kahn, Mark D. LeBlanc & Leah Smith[53]
Wheaton College (Mass.)

[53] Corresponding author: mdrout@wheatoncollege.edu.

Appendix A: Segment Ranges and Contents in *Or*

Segment No.	Word Range	Book & Chapter	Pages & lines in Bately
1	1–900	I.i–I.i	8.11–11.6
2	901–1800	I.i–I.i	11.6–13.23
3	1801–2700	I.i–I.i	13.24–12.2
4	2701–3600	I.i–I.i	16.3–18.6
5	3601–4500	I.i–I.i	18.6–20.20
6	4501–5400	I.i–I.ii–I.iii	20.20–23.3
7	5401–6300	I.iii–I.iv, v, vi, vii.	23.3–25.24
8	6301–7200	I.vii–I.viii	25.24–28.8
9	7201–8100	I.viii–I.ix, x	28.8–30.31
10	8101–9000	I.x–I.xi, xii	30.31–33.21
11	9001–9900	I.xii, I.xiii, I.xiiii–II.I	33.21–36.12
12	9901–10800	II.i, II.ii,	36.12–39.3
13	10801–11700	II.iii–II.iiii	39.3–41.23
14	11701–12600	II.iiii	41.23–44.9
15	12601–13500	II.iiii–II.v	44.9–46.23
16	13501–14400	II.v	46.23–48.35
17	14401–15300	II.v, II.vi, II.vii, II.viii	48.35–51.19
18	15301–16200	II.viii–III.I	51.19–54.4
19	16201–17100	III.i, III.ii, III.iii	54.4–56.24
20	17100–18000	III.iii, III.iv, III.v	56.24–59.24
21	18001–18900	III.v, III.vi, III.vii	59.24–62.14
22	18901–19800	III.vii–III.viii	62.14–64.29
23	19801–20700	III.viii	64.29–67.15
24	20701–21600	III.viii, III.viiii	67.15–70.1
25	21601–22500	III.viiii	70.1–72.25
26	22501–23400	III.viiii–III.x	72.25–75.11
27	23401–24300	III.x–III.xi	75.11–78.1
28	24301–25200	III.xi	78.1–80.22
29	25201–26100	III.xi–IV.i	80.22–83.10
30	26101–27000	IV.i	83.10–85.27
31	27001–27900	IV.i, IV.ii, IV.iii, IV.iiii	85.27–88.16
32	27901–28800	IV.iii–IV.v	88.16–91.6
33	28801–29700	IV.v–IV.vi	91.6–93.29
34	29701–30600	IV.vi	93.29–96.14
35	30601–31500	IV.vi–IV.vii	96.14–99.1
36	31501–32400	IV.vii, IV.viii, IV.ix	99.2–101.19

37	32401–33300	IV.ix–IV.x	101.19–104.7
38	33301–34200	IV.x	104.7–106.25
39	34201–35100	Iv.x–IV.xi	106.25–109.10
40	35101–36000	IV.xi, IV.xii, IV.xiii	109.10–112.3
41	36001–36900	IV.xiii, V.i, V.ii	112.3–114.24
42	36901–37800	V.ii, V.iii	114.24–117.18
43	37801–38700	V.iii, V.iiii, V.v, V.vi, V.vii	117.18–120.20
44	38701–39600	V.vii, V.viii, V.ix, V.x	120.20–123.16
45	39601–40500	V.x., V.xi, V.xii	123.16–126.14
46	40501–41400	V.xii, V.xiii	126.14–129.6
47	41401–42300	V.xiii, V.xiiii, V.xv	129.6–132.2
48	42301–43200	V.xv, VI.i, V.ii	132.2–134.29
49	43201–44100	V.ii, VI.iii, VI.v	134.29–137.23
50	44101–45000	VI.v, VI.vi, VI.vii, VI.viii., VI.viiii, VI.x, VI.xi, VI.xii, VI.xiii	137.23–141.7
51	45001–45900	VI.xiii, VI.xiiii, VI.xv, VI.xvi, VI.xvii, VI.xviii, VI.xviiii, VI.xx, VI.xxi, VI.xxii, VI.xxiii	141.7–144.18
52	45901–46800	VI.xxiii, VI.xxiiii, VI.xxv, VI.xxvi, VI.xxvii, VI.xxviii, VI.xxviiii, VI.xxx	144.18–148.7
53	46801–47700	VI.xxx, VI.xxxii	148.7–151.6
54	47701–48600	VI.xxxii, VI.xxxiii, VI.xxxiiii, VI.xxxv, VI.xxxvi	151.6–154.4
55	48601–49452	VI.xxxvi, VI.xxxvii, VI.xxxviii	154.4–156.23

References

Bately, J. 1970: King Alfred and the Old English Translation of Orosius. *Anglia* 88: 433–460.

Bately, J. 1971: The Classical Editions in the Old English Orosius. In P. Clemoes & K. Hughes eds. *England Before the Conquest*. Cambridge, Cambridge University Press: 237–251.

Bately, J. ed. 1980: *The Old English Orosius* (E.E.T.S. S.S. 6). London, Oxford University Press.

Burrows, J. F. 2003: Questions of Authorship: Attribution and Beyond. *Computers and the Humanities* 37: 5–32.

Cameron, A. & R. Frank 1973: *A Plan for the Dictionary of Old English*. Toronto, University of Toronto Press.

Campbell, J. 1959: *Old English Grammar*. Oxford, Clarendon Press.

Cerquiglini, B. 1999: *In Praise of the Variant: A Critical History of Philology*. [Betsy Wing trans. 1989: *Éloge de la variante*]. Baltimore, Johns Hopkins University Press.

Chauvet, E. & M. D. C. Drout forthcoming: Visual Representation of the Ratio of þ to þ+ð: A New Tool for the Investigation of Old English Textual History.

Crick, J. 1987: An Anglo-Saxon fragment of Justinus' *Epitome*. *Anglo-Saxon England* 16: 181–196.

Derolez, R. 1971: The orientation system in the Old English Orosius. In P. Clemoes & K. Hughes eds. *English Before the Conquest*. Cambridge, Cambridge University Press: 253–268.

Downey, S., M. D. C. Drout, M. Kahn & M. LeBlanc 2012: 'Books Tell Us': Lexomic and Traditional Evidence for the Sources of *Guthlac A*. *Modern Philology* 110: 1–29.

Downey, S., M. D. C. Drout, V. Kerekes & D. Raffel [forthcoming]: Lexomic Analysis of Medieval Latin Texts.

Drout, M. D. C. 2013: *Tradition and Influence in Anglo-Saxon Literature: An Evolutionary, Cognitivist Approach*. New York, Palgrave Macmillan.

Drout M. D. C. & S. Kleinman 2010: Philological Inquiries 2: Something Old, Something New: Material Philology and the Recovery of the Past. *The Heroic Age* 13. http://www.mun.ca/mst/heroicage/issues/13/pi.php (accessed 2 March 2013).

Drout, M. D. C., M. Kahn, M. LeBlanc & C. Nelson 2011: Of Dendrogrammatology: Lexomic Methods for Analyzing the Relationships Among Old English Poems. *Journal of English and Germanic Philology* 110: 301–336.

Drout, M. D. C., Y. Kisor, A. Dennett, N. Piirainen & L. Smith forthcoming: Lexomic Analysis of *Beowulf*.

Dyer, B. 2002: *Genome Technology* 1.27. http://www.genomeweb.com/blunt-end-0. (accessed 1 November 2002).

Frantzen, A. J. 1983: *The Literature of Penance in Anglo-Saxon England*. New Brunswick (NJ), Rutgers University Press.

Frantzen, A. J. 2013: *The Anglo-Saxon Penitentials: A Cultural Database*. http://www.anglo-saxon.net/penance (accessed 2 March 2013).

Gneuss, H. 2001: *Handlist of Anglo-Saxon Manuscripts*. Tempe (AZ), Arizona Medieval and Renaissance Texts and Studies.

Hennig, W. 1966: *Phylogenetic Systematics* [D. D. Davis & R. Zangerl trans. 1950: *Grundzüge einer Theorie der phylogenetischen Systematik*]. Urbana, University of Illinois Press.

Hoover, D. L. 2004: Testing Burrows's Delta. *Literary and Linguistic Computing* 19.4: 453–475.

Ker, N. R. 1957: *Catalogue of Manuscripts Containing Anglo-Saxon*. Oxford, Clarendon Press.

Keynes, S. & M. Lapidge 1983: *Alfred the Great: Asser's Life of King Alfred and Other Contemporary Sources*. London, Penguin.

Lapidge, M. 2003: Asser's Reading. In T. Reuter ed. *Alfred the Great*. London, Ashgate.

Liggins, E. 1970: The Authorship of the Old English Orosius. *Anglia* 88: 289–322.

Mardia, K, J. Kent & J. Bibby 1980: *Multivariate Analysis*. London, Academic Press.

Megginson, D. 1993: *The Written Language of Old English Poetry*. (PhD Dissertation). Toronto, University of Toronto.

Millett, B. 2008: What is *mouvance*? http://www.soton.ac.uk/~wpwt/mouvance/mouvance.htm (accessed 12 Dec 2012).

O'Brien O'Keeffe, K. 1990: *Visible Song: Transitional Literacy in Old English Verse*. Cambridge, Cambridge University Press.

O'Donnell, D. P. 2005: *Cædmon's Hymn: A Multimedia Study, Edition and Archive*. Woodbridge, D. S. Brewer.

R Development Core Team 2009: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. http://www.R-project.org (accessed 2 March 2013).

Raith, J. ed. 1964 [1933]: *Die altenglische Version des Halitgar'schen Bussbuches (sog. Poenitentiale Pseudo-Ecgberti)*. Darmstadt, Wissenschaftliche Buchgesellschaft.

Raith, J. 1951: *Untersuchungen zum englischen Aspekt, I. Grundsätzliches Altenglisch*. Munich, Heuber.

Roberts, J. 2006: *Guide to Scripts Used in English Writings up to 1500*. London, British Library.

Schmitz, H. J. ed. 1958 [1898]: *Die Bussbücher und das kanonische Bussverfahren*. Graz, Akademische Druck U. Verlaganstalt.

Schröer, A. ed. 1964 [1885]: *Die Angelsächsischen Prosabearbeitungen der Benediktinerregel*. Darmstadt Wissenschaftliche Buchgesellschaft.

Seel, O. 1972: *M. Iuniani Iustini epitoma Historiarum Philippicarum Pompei Trogi*. Stuttgart, B. G. Teubner.

Shippey, T. 2007: Fighting the Long Defeat: Philology in Tolkien's Life and Works. *Roots and Branches: Selected Papers on Tolkien by Tom Shippey*. Jena, Walking Tree Publishers.

Shippey, T. 2008: Response to three papers on 'Philology: Whence and Whither?' given by Drs Utz, Macgillivray, and Zolkowski, at Kalamazoo, 4th May 2002. *The Heroic Age* 11: http://www.mun.ca/mst/heroicage/issues/11/foruma.php (accessed 2 March 2013).

Spindler, R. ed. 1934: *Das altenglische Bussbuch (sog. Confessionale Pseudo-Egberti)*. Leipzig, Tauchnitz.

Stokes, P. 2009: The Digital Dictionary. *Florilegium* 26: 37–65.

Stubbs, W. ed. 1887: William of Malmesbury *De Gestis Regis Anglorum* (Rolls Series 90). London, Longman.

Sweet, H. 1883: *King Alfred's Orosius: Part I: Old English Text and Latin Original* (E.E.T.S. O.S. 79). London, Trübner.

Thorpe, B. 1840: *Ancient Laws and Institutes of England*. 2 vols. London, G. E. Eyre and A. Spottiswoode.

Townend, M. 2002: *Language and History in Viking Age England: Linguistic Relations between Speakers of Old Norse and Old English*. Brepols, Turnhout.

Valtonen, I. 2008: *The North in the Old English Orosius. A Geographical Narrative in Context*. Helsinki, Société Néophilologique.

Zumthor, P. 1972: *Essai de poétique médiévale*. Paris, Seuil.

Zumthor, P. 1987: *La lettre et la voix: de la 'littérature' médiévale*. Paris, Seuil.

Received 16 Apr 2013; *accepted* 14 Sep 2013