Lexomics Analysis and the Problem of Editions

(Slide 1)

(Slide 2)

Lexomic research forces us to think very carefully about the editions we are using. Editorial decisions about orthography, word division, and emendation have the potential to affect significantly the data we are using to produce dendrograms and analyze textual structures and relationships. Texts from before the age of mechanical reproduction are particularly influenced by editorial practices.
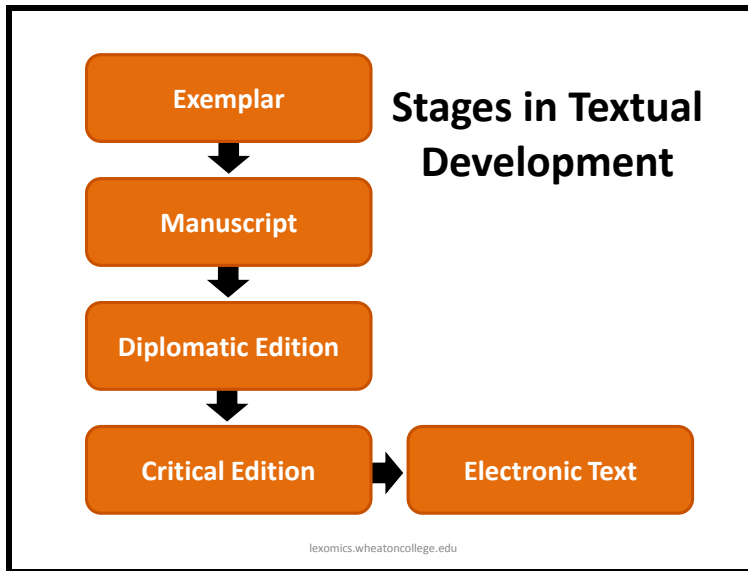
(Slide 3)

This chart provides a simplified representation of the stages of textual development.
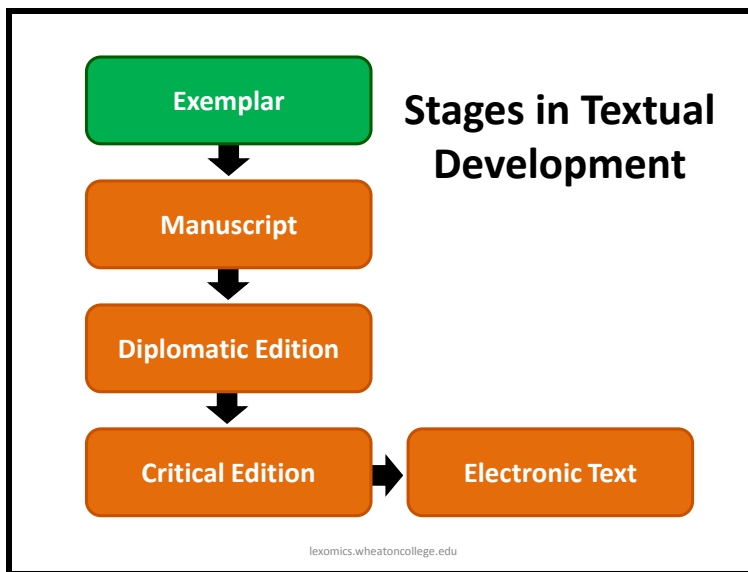
(Slide 4)

An exemplar is a text from which a later manuscript is copied. Eventually all exemplars go back to the original version of a text.
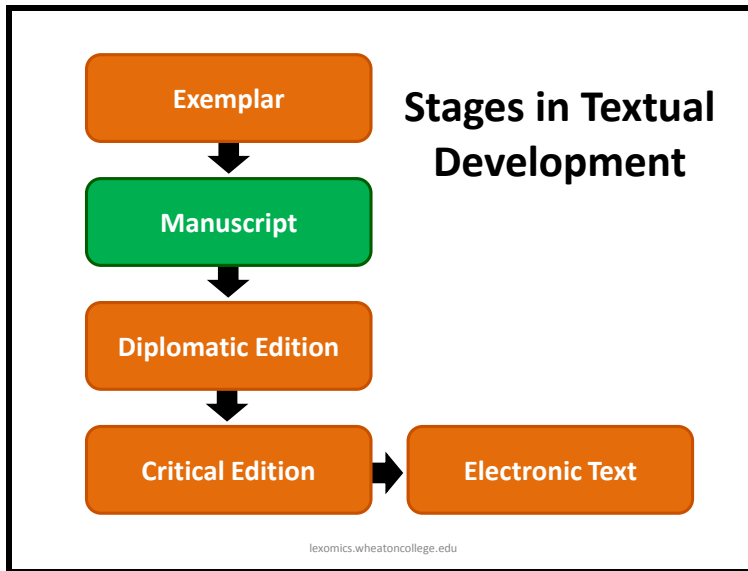
(Slide 5)

Most exemplars are lost to us (though some autograph copies of very old texts to exist). The exemplar was copied to produce a manuscript, a handwritten document that has been preserved.
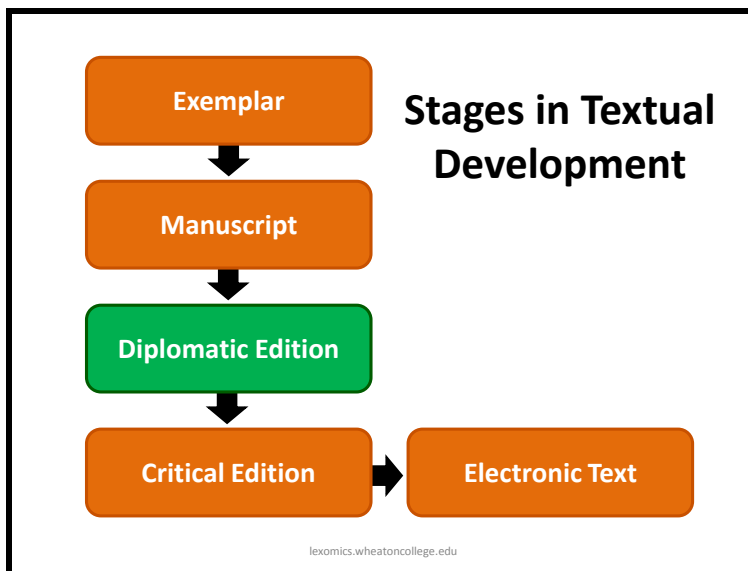
(Slide 6)

Editors begin with the manuscripts that still exist and simply transcribe these—copying exactly what text is on the page—to produce a print copy called a diplomatic edition. The diplomatic edition does not add anything to the text, even reproducing errors and omissions.
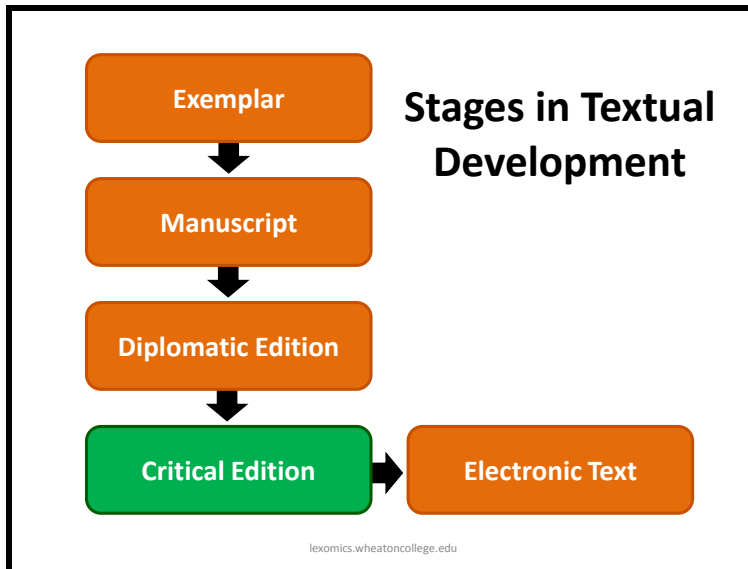
Applying their philological, historical and culture knowledge, editors then produce a critical edition from the diplomatic one. A critical edition corrects errors, fills in blanks and produces a version of the text that is—at least in the mind of the editor—as close to what the original author produced as is possible.
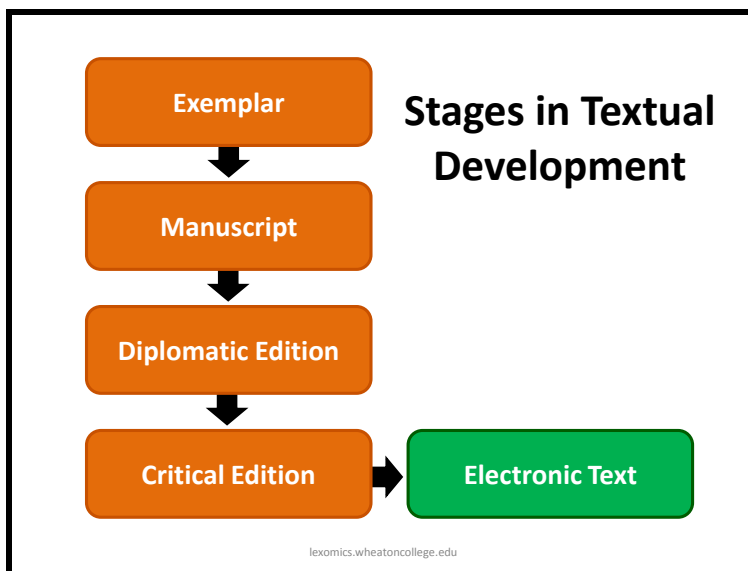
Almost all electronic texts are then produced from critical editions, often by scanning the critical edition and then correcting the electronic text against the print version, sometimes—especially in more recent years—by using the editor's own electronic version of the critical edition.

The Anglo-Saxon *Genesis*

lexomics.wheatoncollege.edu

(Slide 9)

Let's look at the stages of textual development for the Anglo-Saxon poem Genesis.

Junius Manuscript (*Genesis*)

Diplomatic Edition

ASPR Critical Edition

DOE Electronic Text

Textual Development of *Genesis*

lexomics.wheatoncollege.edu
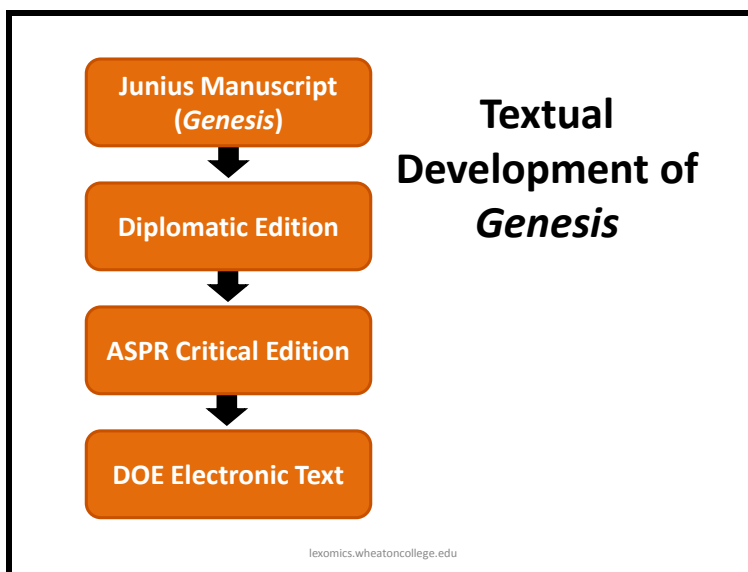
(Slide 10)

Some time in the tenth century an exemplar was copied by a scribe to produce the manuscript now known as Oxford, Bodleian Library, Junius xi. This manuscript was transcribed by the great scholar Francis Junius, and also later by other scholars to produce personal diplomatic editions. From these, later scholars produced critical editions. The most widely used critical edition is by George Philip Krapp, part of the Anglo-Saxon Poetic Records (ASPR) series published by Columbia University Press.

The Dictionary of Old English produced an electronic version of Krapp's text for use in the Dictionary of Old English Corpus. We have used that electronic edition for Lexomic analysis.
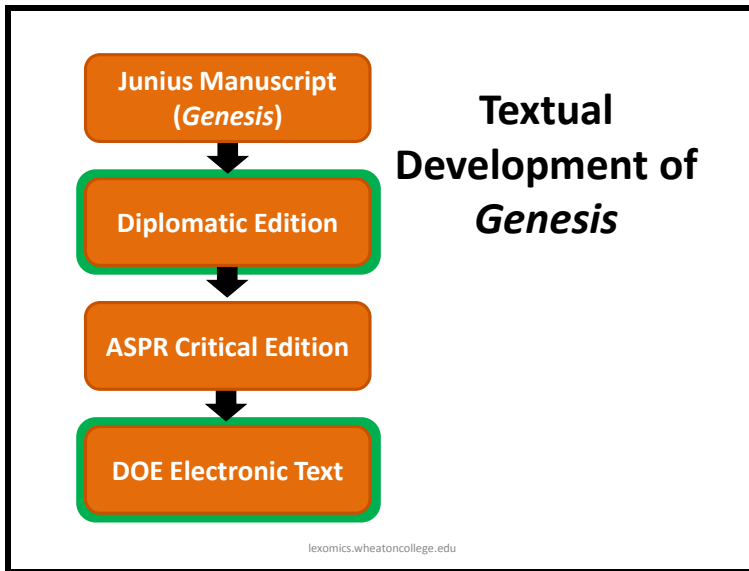
(Slide11)

To see the effects of editorial decisions on our Lexomic analysis, we can compare the results we get from analyzing an electronic *diplomatic* edition with results from analysis of an *electronic* critical edition.

Unfortunately, there are very few electronic diplomatic editions of Old English texts (though editors are beginning to produce them). There was not, at the time we did this research, an electronic diplomatic edition of *Genesis*, so we needed to create one.
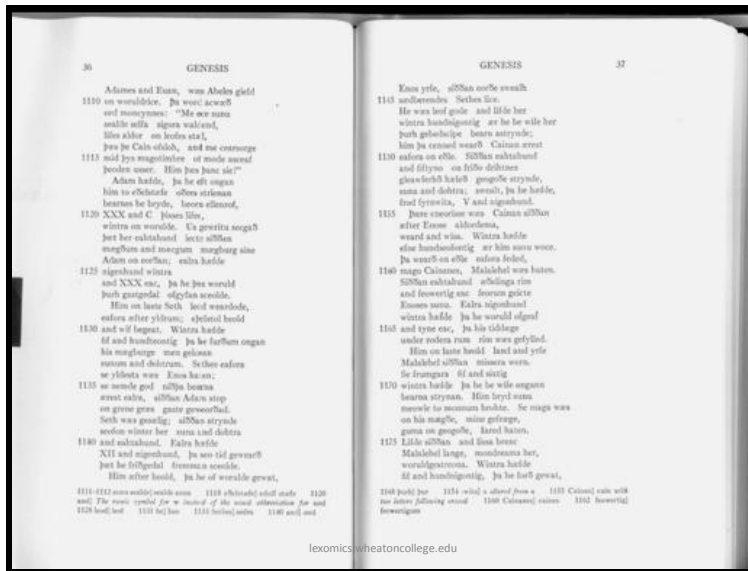
(Slide 12)

Here is a page from Krapp's critical edition of *Genesis* in the Anglo-Saxon poetic records.

(Slide 13)

You can see the emendations in the footnotes. These record the forms in the original manuscript.

(Slide 14)

Here, in line 1128, the manuscript has the word "leof,"
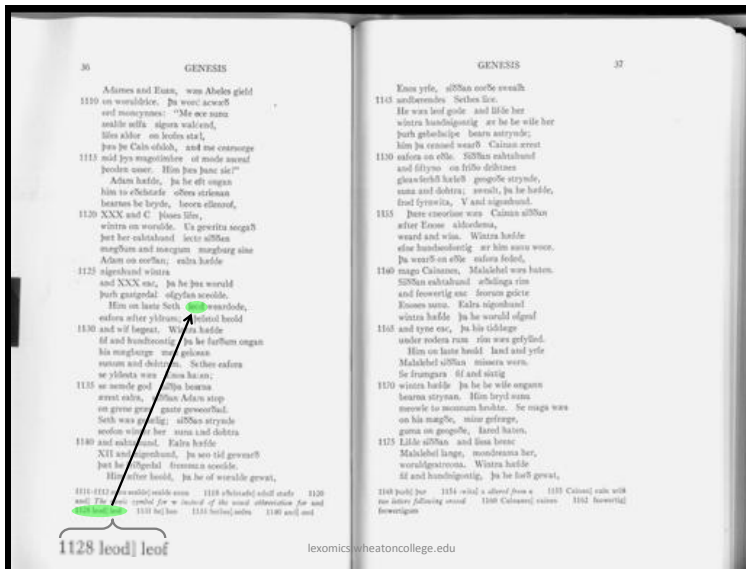
(Slide 15)

...which the editors changed to "leod" in the critical edition in order to have the text make more sense at this point.

By taking the electronic file of the critical text and using the information in the notes to change all the critical edition forms to the manuscript forms, our research partner Scott Kleinman was able to create an electronic diplomatic edition of *Genesis*.

Dendrogram of Diplomatic Edition of *Genesis AB (1500-word chunks)*
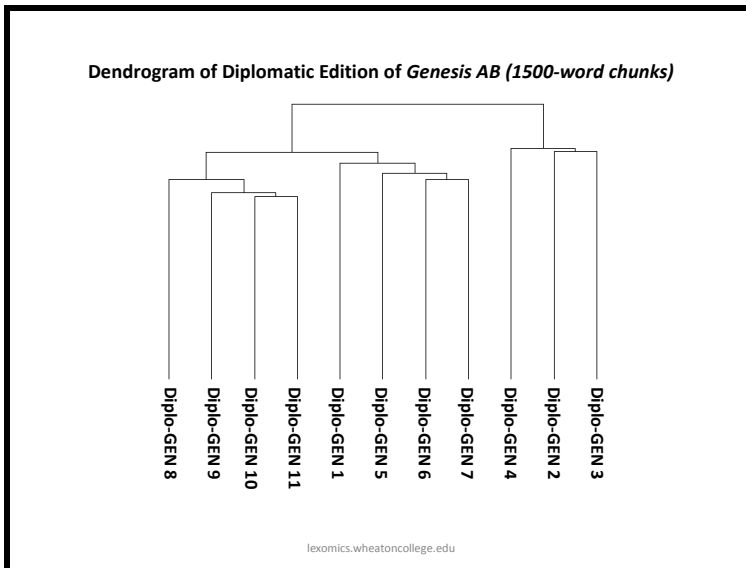
(Slide 16)

We then used our software to produce a dendrogram of the distribution of vocabulary in this diplomatic version of *Genesis.*
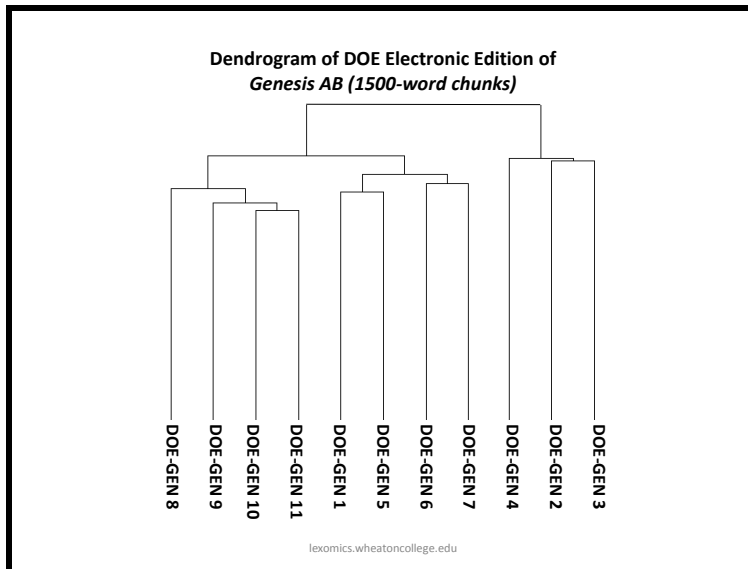
(Slide 17)

For purposes of comparison we then created a dendrogram from the critical edition using the Dictionary of Old English electronic text of *Genesis*.
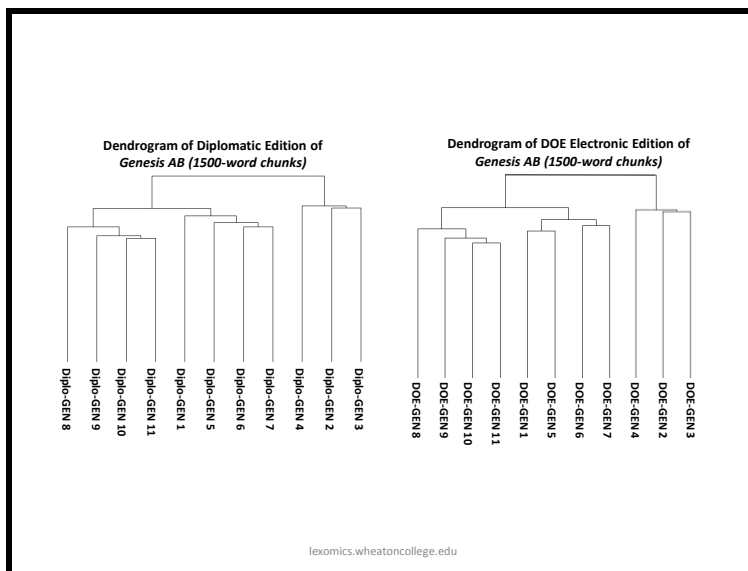
(Slide 18)

Putting these dendrograms side by side allows us to see large similarities and minor differences. Note that both have identical high-level clade structures. The first division separates *Genesis B* from *Genesis A*. 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.
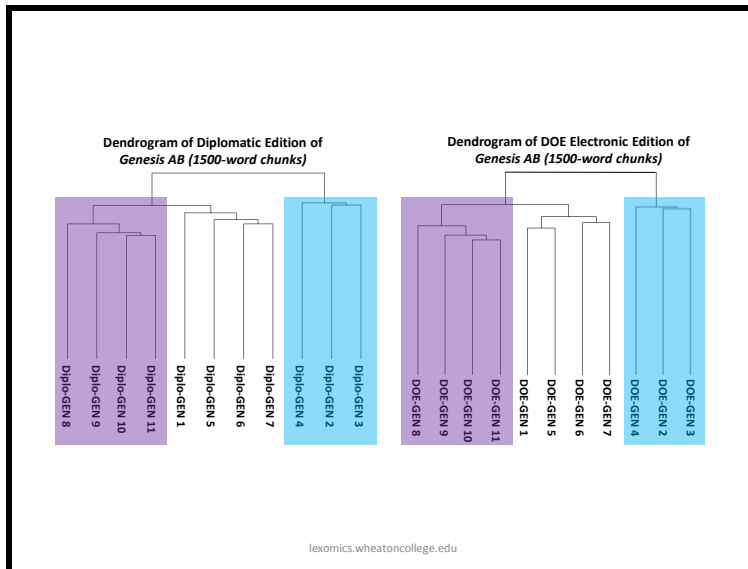
(Slide 19)

Moving deeper into the clade structure, we see additional similarities. The clades highlighted in purple and blue on the dendrograms have exactly the same geometries in the diplomatic and critical editions.
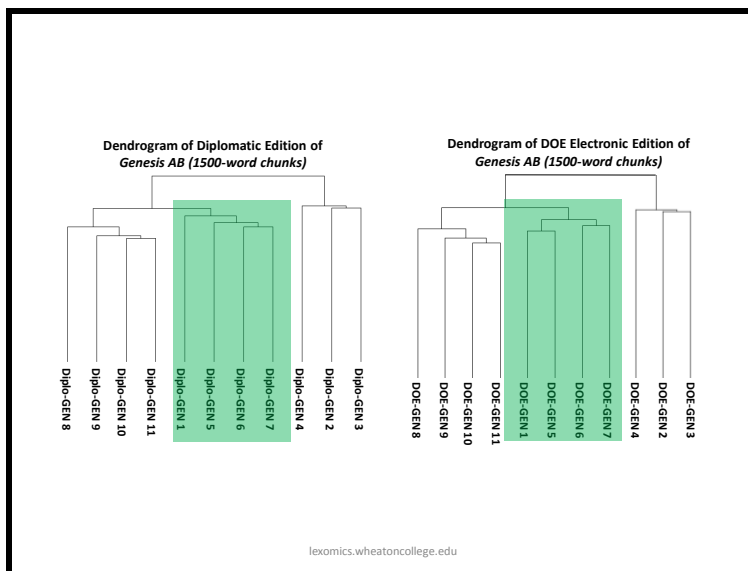
(Slide 20)

Only the middle clade, highlighted here in green, shows some minor variation. In both texts chunks 6 and 7 are the most similar, but in the critical edition chunks 5 and 1 are paired, while in the diplomatic edition they join the 6-7 clade in a stepwise fashion. This geometry indicates that the diplomatic edition is slightly more variable than the critical edition—exactly what we would expect if we realize that the diplomatic edition contains more errors than the critical edition. The close similarity of the rest of the geometry of the dendrograms indicates that there is really very little difference between the critical edition and the diplomatic edition. Therefore the critical edition can be used as long as we are not drawing conclusion based solely on very subtle differences in geometry deep within clades.

Daniel and *Azarias*

lexomics.wheatoncollege.edu

(Slide 21)

The poems *Daniel* and *Azarias* allow us to look at a different kind of relationship than that between *Genesis* and *Genesis B*. *Azarias* is quite similar to a section of *Daniel* because both derive from the same source: an older Anglo-Saxon poetic translation of the biblical book of *Daniel*.

Junius Manuscript (*Daniel*) → Diplomatic Edition → ASPR Critical Edition → DOE Electronic Text

The Exeter Book (Azarias) → Diplomatic Edition → ASPR Critical Edition → DOE Electronic Text

lexomics.wheatoncollege.edu
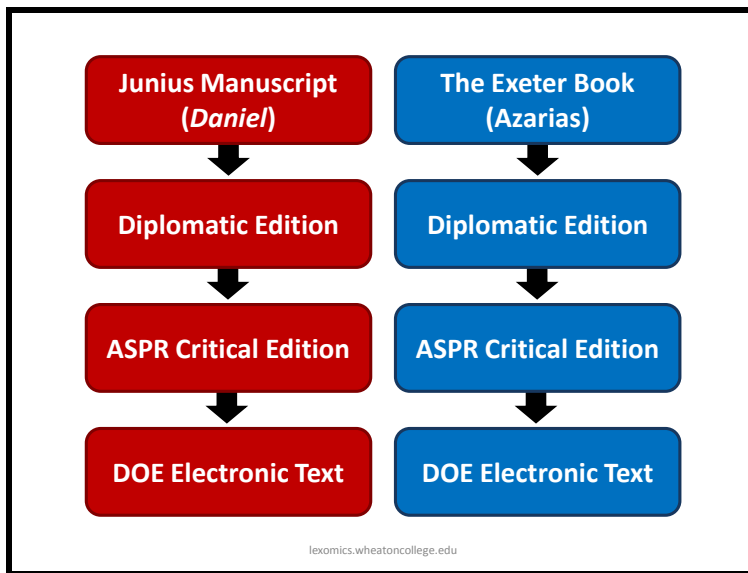
(Slide 22)

Daniel and *Azarias* were copied into different manuscripts, the Junius Manuscript and Exeter Book. The editors of the Anglo-Saxon Poetic Records created critical editions of the poems, and these editions were used for the Dictionary of Old English electronic corpus. Scott Kleinman reconstructed diplomatic editions of the poems by using the critical apparatus to emend the electronic files so that they match the text in the manuscript rather than in the critical edition.
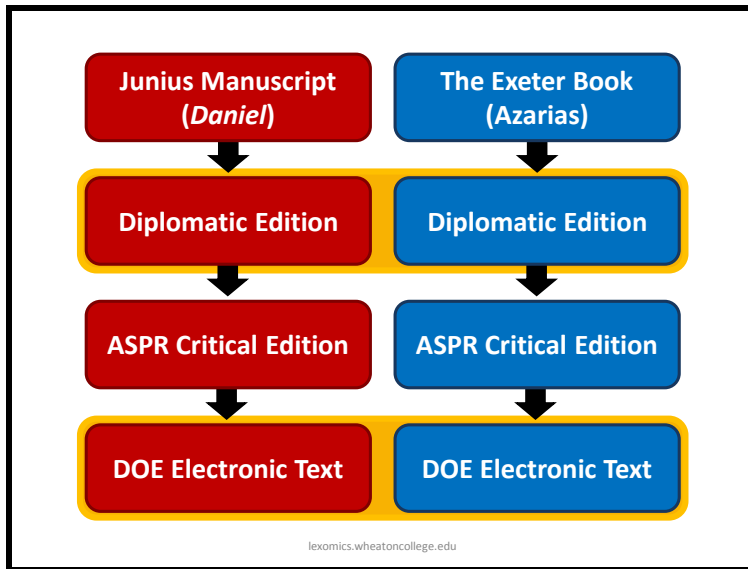
(Slide 23)

As we did in the *Genesis* experiment, we compared the electronic Dictionary of Old English texts to the diplomatic editions of both texts. Here is the dendrogram of the diplomatic editions of *Daniel* and *Azarias*.
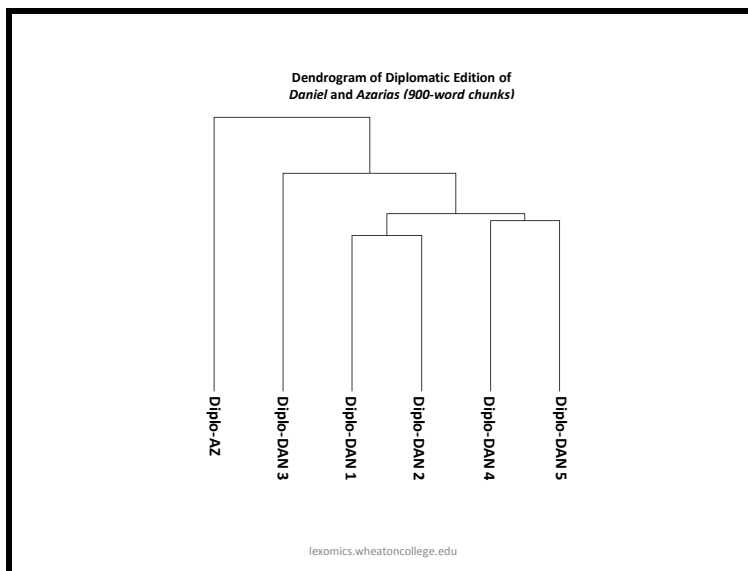
(Slide 24)

And here is the dendrogram of the Dictionary of Old English critical editions of *Daniel* and *Azarias*.
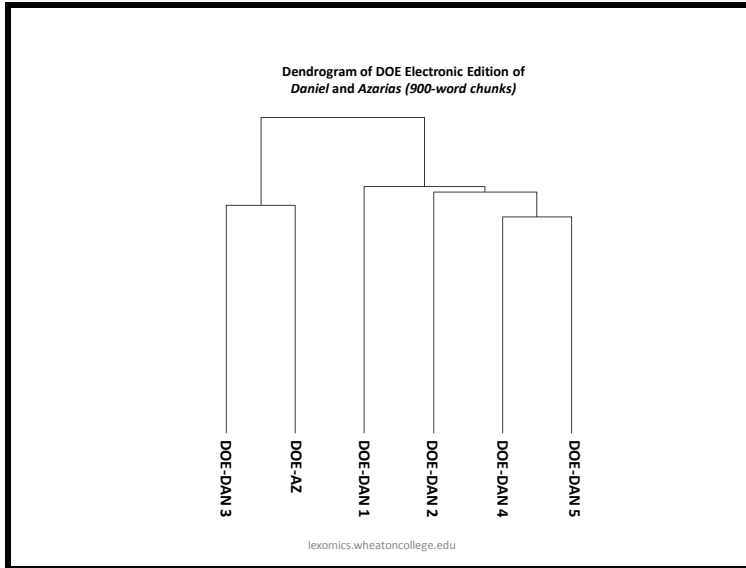
(Slide 25)
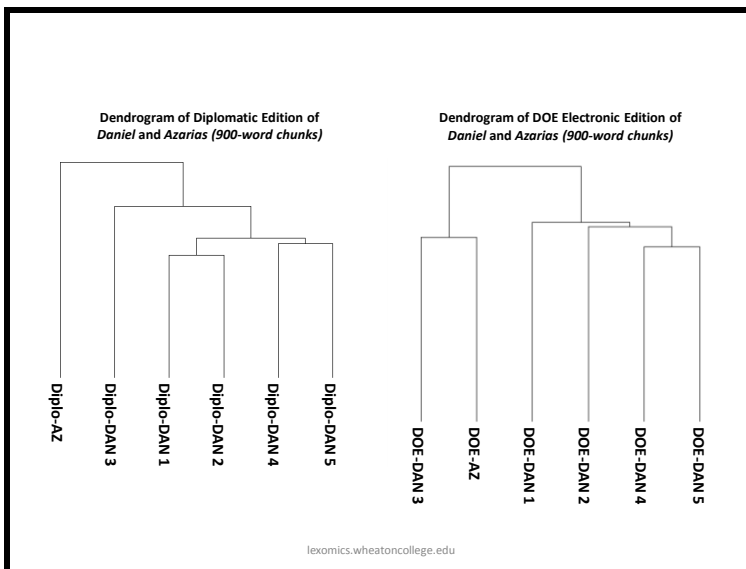
(Slide 26)

Comparing the dendrograms of the diplomatic and critical editions, we see that both separate *Azarias* from *Daniel*, and both correctly identify the third 900-word segment of *Daniel* (Dan 3) as being most similar to *Azarias.*

Dendrogram of Diplomatic Edition of *Daniel* and *Azarias (900-word chunks)*

Dendrogram of DOE Electronic Edition of *Daniel* and *Azarias (900-word chunks)*

lexomics.wheatoncollege.edu

(Slide 27)

The larger clade structure is essentially the same: the first, second, fourth and fifth chunks of *Daniel* are similar to each other, and *Azarias* is an outlier along with the third chunk of *Daniel*. The differences in the two dendrograms are found deeper inside the clades. In the dendrogram created from the diplomatic edition, the *Azarias* and *Daniel 3* clades 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 chunks 1 and 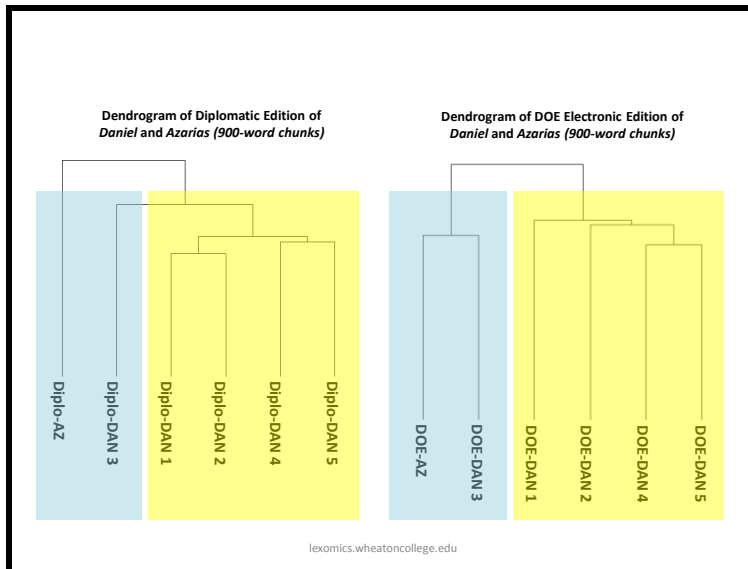2 are paired in the diplomatic edition but are slightly separated in the critical edition. Based on what we know of *Daniel* and *Azarias* from traditional methods—including simply reading the poems and comparing them line-by-line and word-by-word—we can conclude that the dendrogram created from the critical edition is more consistent with the actual relationship between the two poems. *Azarias* is very much like the third chunk of *Daniel*, and both of these are less like the rest of the poem. We therefore conclude that although there is not much difference between the two editions, it is probably preferable to use the critical edition for Lexomic analysis.

Editions Collated from Multiple Manuscripts

lexomics.wheatoncollege.edu
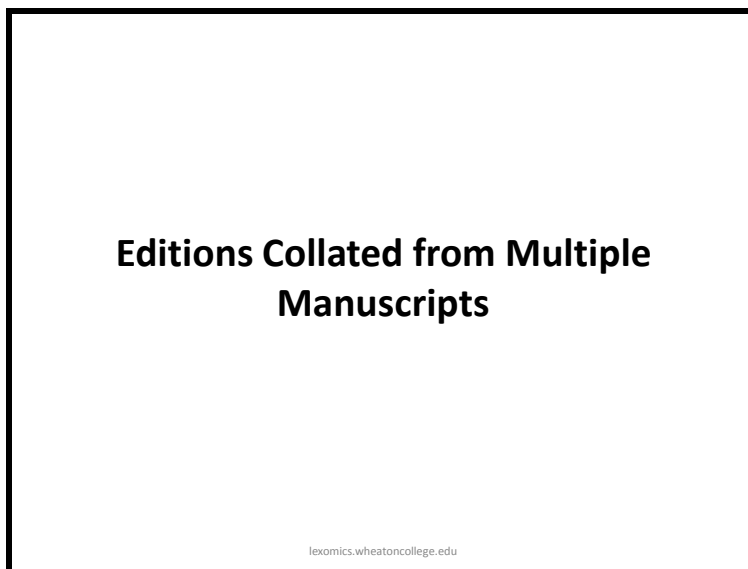
(Slide 28)

Most Anglo-Saxon poems are found in only one manuscript. But many Anglo-Saxon prose texts, as well as many medieval texts in other languages, have multiple manuscript witnesses.
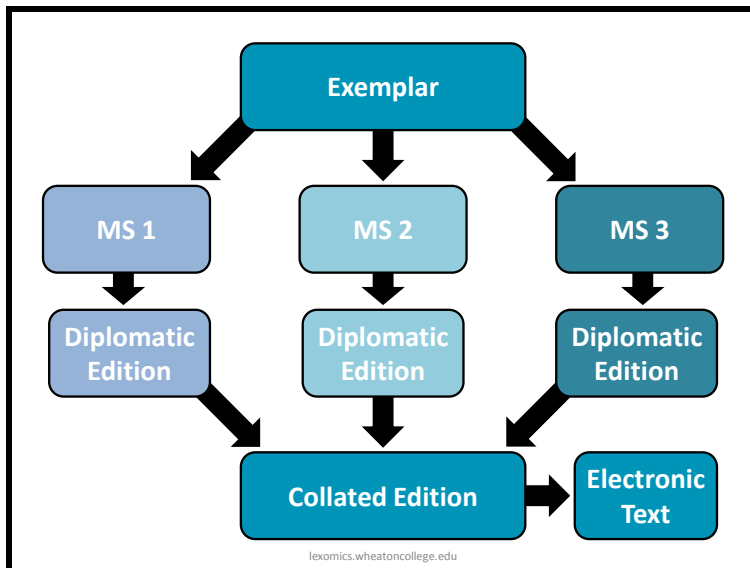
The editions produced from these texts have a somewhat more complicated textual history. If we go back far enough in a textual tradition, we can find a single exemplar from which multiple copies were made, each of which differs from the other to greater or lesser degrees. Editors therefore produce transcripts or diplomatic editions of each of the manuscript witnesses, and then they collate and compare them to produce a single critical edition, from which electronic texts are produced. A difficulty with critical editions created in this way is that they at times will blend together features of multiple texts, producing a work that is assumed to be as close to the author's original as possible, but which, if the editor's reconstructions are incorrect, may represent a text that never existed.

The *Old English Penitential*

lexomics.wheatoncollege.edu

The Old English Penitential is a text with multiple manuscript witnesses. These were collated to produce the traditional critical edition that is used by the Dictionary of Old English. But thanks to Allen Frantzen's work, we have electronic versions of each of the major manuscripts. We can therefore compare the dendrograms produced from these texts to see if or how they are different from dendrograms produced from the collated edition used by the Dictionary of Old English.

The *Old English Penitential*
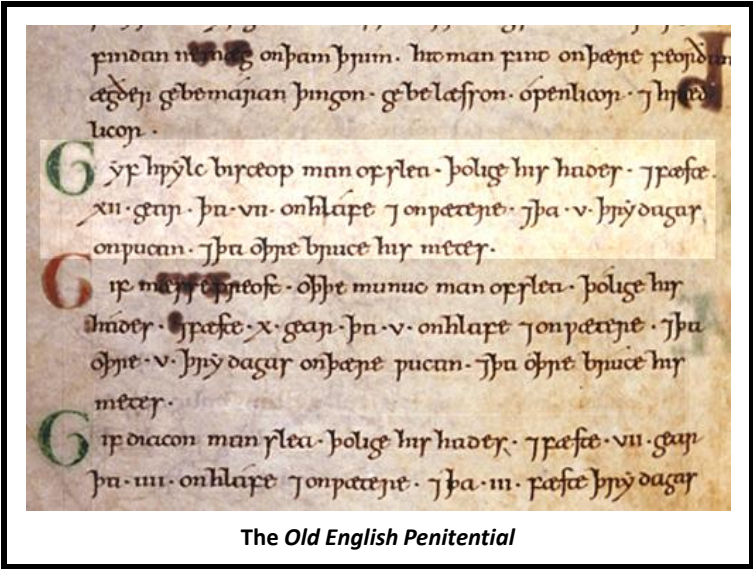
(Slide 31)

A penitential is a list of sins and the appropriate penances that people who had committed the sins could use to cleanse their souls. Penitentials were invented in Ireland, traveled to England, crossed over to continental Europe, and then were re-imported to England. Most penitentials were originally written in Latin and only later translated into the vernacular.
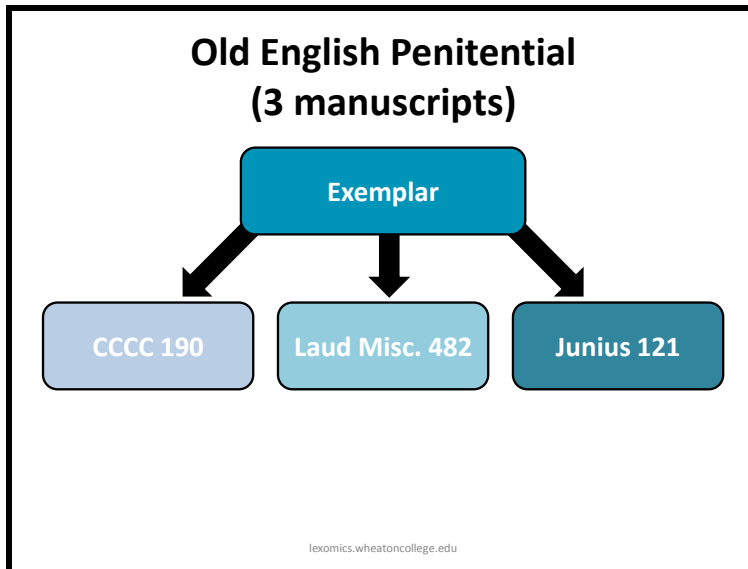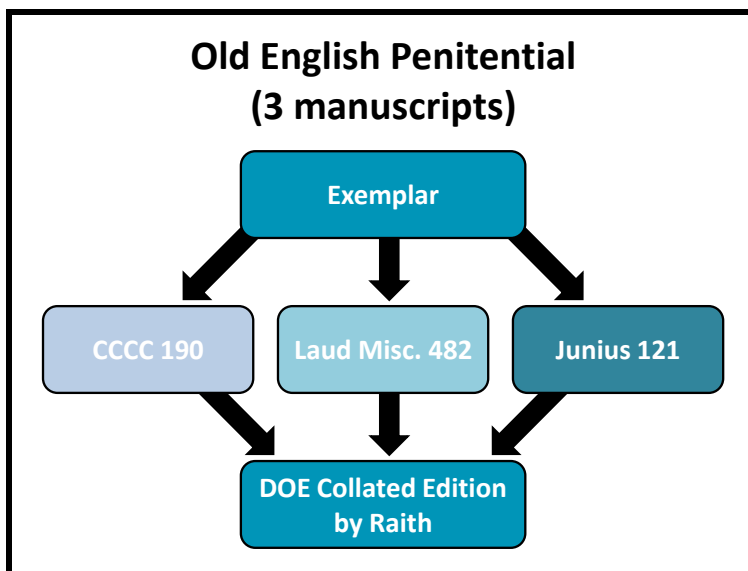

**Old English Penitential
(3 manuscripts)**

Exemplar

lexomics.wheatoncollege.edu

(Slide 32)

Old English Penitential
(3 manuscripts)

Exemplar

CCCC 190   Laud Misc. 482   Junius 121

lexomics.wheatoncollege.edu

(Slide 33)

Old English Penitential
(3 manuscripts)

Exemplar

CCCC 190   Laud Misc. 482   Junius 121

DOE Collated Edition by Raith

(Slide 34)

The particular text we are examining, the *Old English Pentitential,* is primarily a translation of a Latin penitential written by Haltigar, bishop of Cambrai, in the ninth century. It is found in four manuscripts, though for the purpose of this discussion we will focus on only three (as the amount of material in the fourth is very small).

The three major manuscripts are: Cambridge, Corpus Christi College 190; Oxford, Bodleian Library Junius 121; and Oxford, Bodleian Library, Laud Misc. 482. These were used by Josef Raith to produce a collated critical edition.
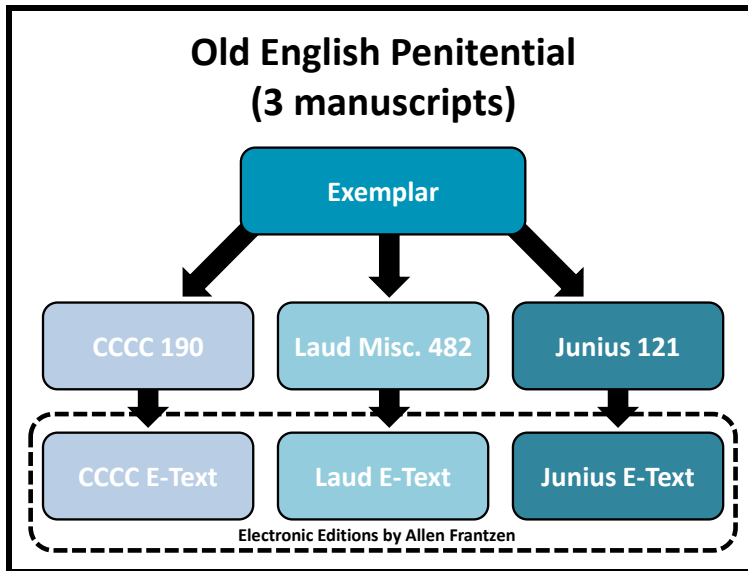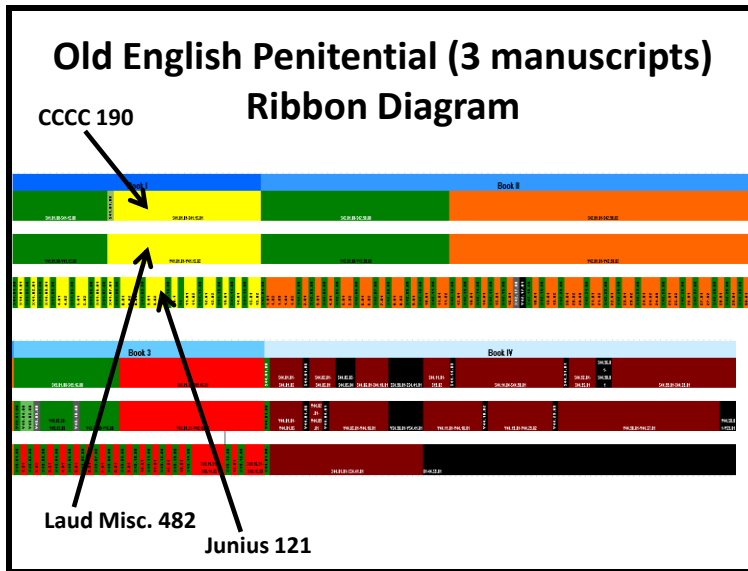
(Slide 35)

Much more recently, Allen J. Frantzen produced electronic editions of each manuscript, which he has allowed us to use for this study.
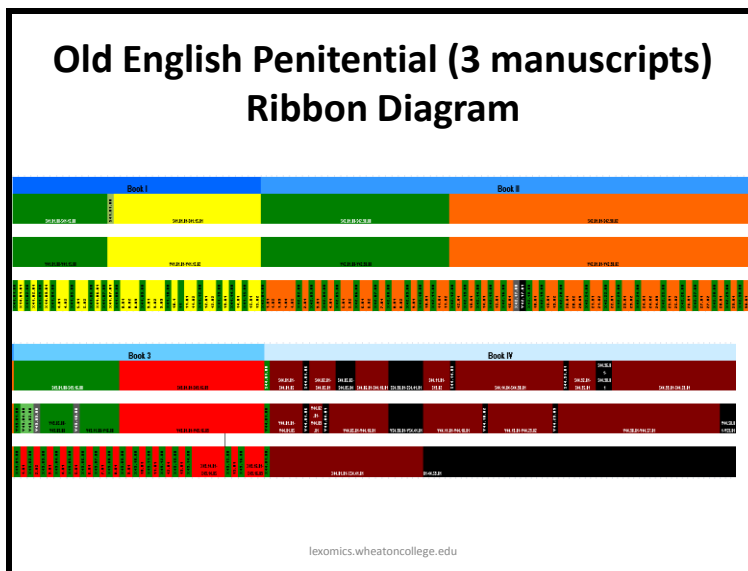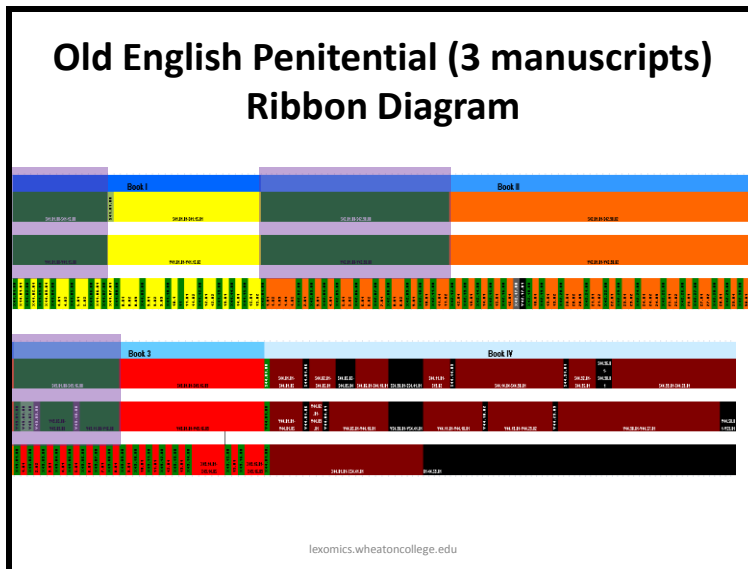

(Slide 36)

To get a sense of how each manuscript is structured, we created a ribbon diagram, dividing the manuscripts based on content. The top blue bar of the ribbon indicates the different books of the penitential.

Old English Penitential (3 manuscripts) Ribbon Diagram
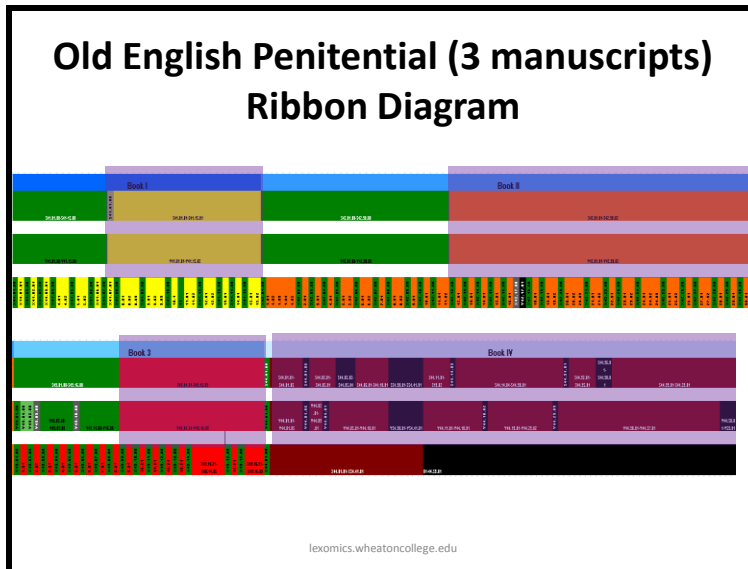
(Slide 37)

The lower bars represent the organization of each manuscript: Corpus 190 is the upper ribbon, Laud Misc. 482 in the middle, and Junius 121 on the bottom.

Old English Penitential (3 manuscripts) Ribbon Diagram

(Slide 38)

Notice that the top two manuscript ribbons match up almost exactly. The green sections, highlighted here, are groups of capitulae, or chapter titles,

Old English Penitential (3 manuscripts) Ribbon Diagram

(Slide 39)

...and the yellow, orange, and red sections, now highlighted, are the text of Books One, Two, and Three, respectively.


Old English Penitential (3 manuscripts) Ribbon Diagram
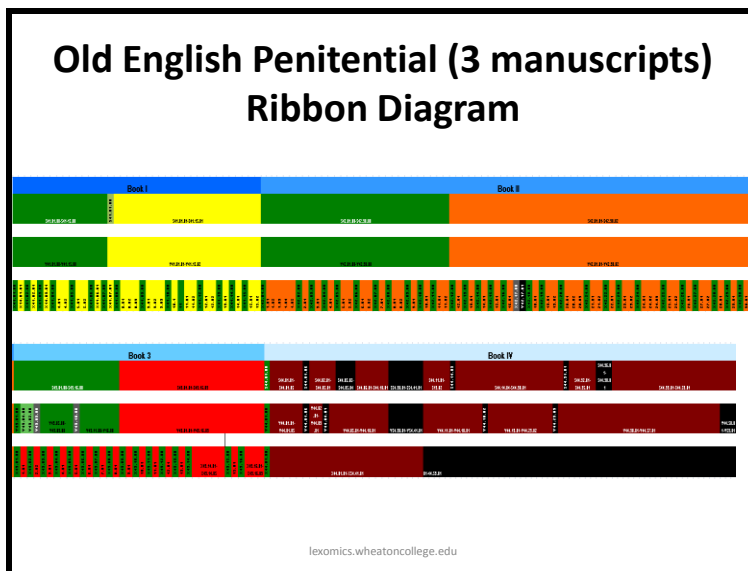
(Slide 40)

Book Four, indicated by the maroon section, is incomplete in all three manuscripts, with the missing portions indicated in black.

Old English Penitential (3 manuscripts) Ribbon Diagram

lexomics.wheatoncollege.edu
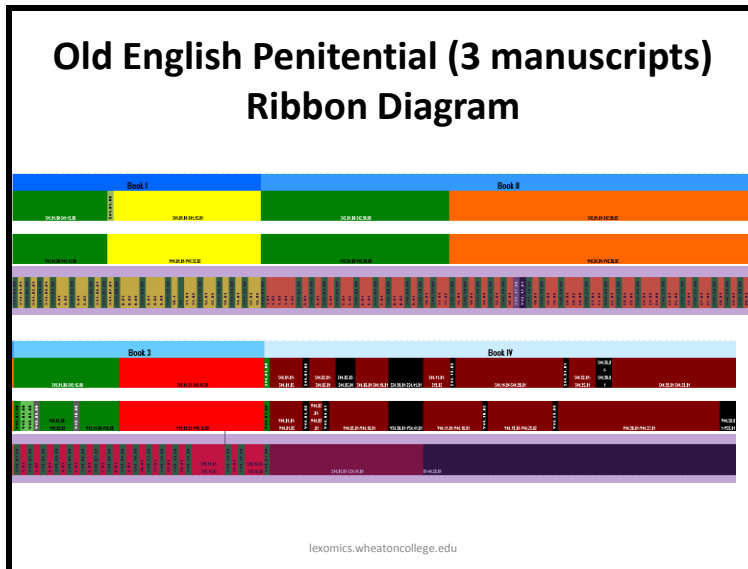
(Slide 41)

In the Junius manuscript, the capitulae, marked in green, are interspersed throughout Books One, Two, and Three. Instead of being a large table of contents, as they are in the Corpus and Laud manuscripts, in Junius they are the headings for individual sections of the text.

Old English Penitential (3 manuscripts) Ribbon Diagram

MISSING

lexomics.wheatoncollege.edu
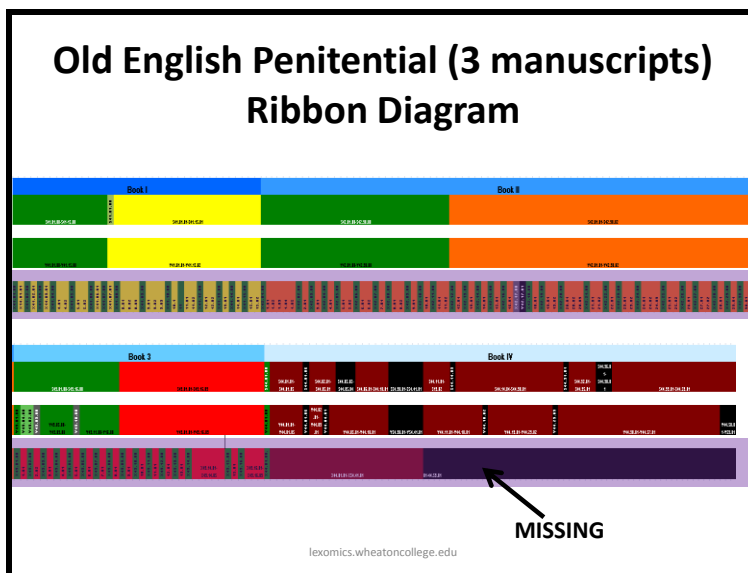
(Slide 42)

The Junius Manuscript is also missing the most material in book four. Book four is taken from the Anglo-Saxon penitential known as the *Scriftboc* and so, unlike Books 1-3, is not based on Haltigar's Latin text.

Having the capitulae spread throughout the Junius text creates some challenges for Lexomic analysis. Although Corpus and Laud match up chunk-for-chunk, the same content is distributed somewhat differently in Junius, making it difficult to compare all three manuscripts even though they have essentially the same content and come from the same source.
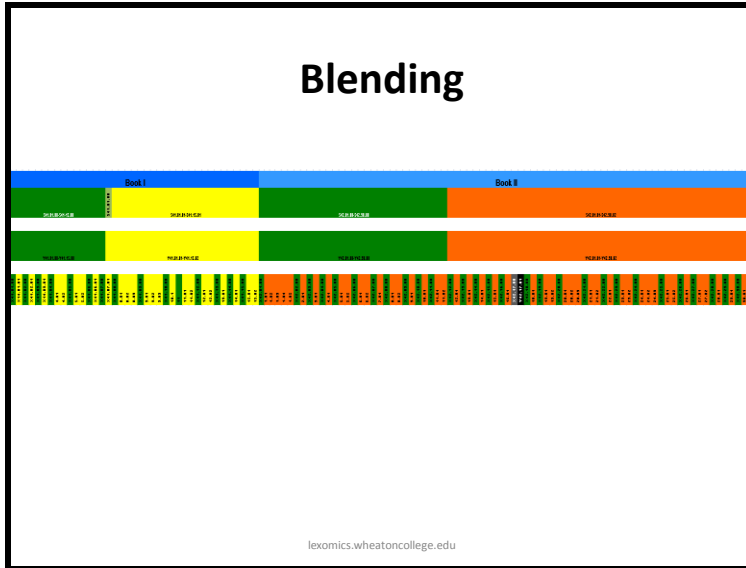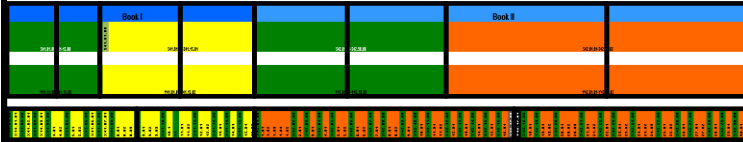
(Slide 43)

To combat this problem, we used a process called blending to modify the Corpuss and Laud manuscripts so that each segment has the same content as the matching segment in the Junius text.
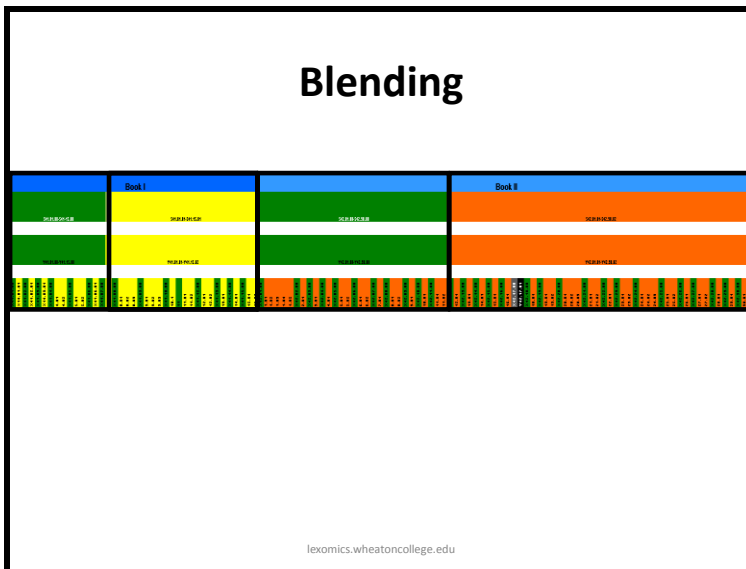
(Slide 44)

Here you see the original chunks across the three manuscripts for Books One and Two. The Corpus and Laud chunks match up with each other: chunk 1 contains the capitulae of Book One, chunk 2 contains the text of Book One, chunk 3 contains the capitulae of Book two, and chunk 4 the text of Book Two. But in Junius, the capitulae are spread throughout the text, therefore, although the first 1000 words of Laud and Corpus and entirely capitulae, the first 1000 words of Junius are approximately 65 percent text and 35 percent capitulae.
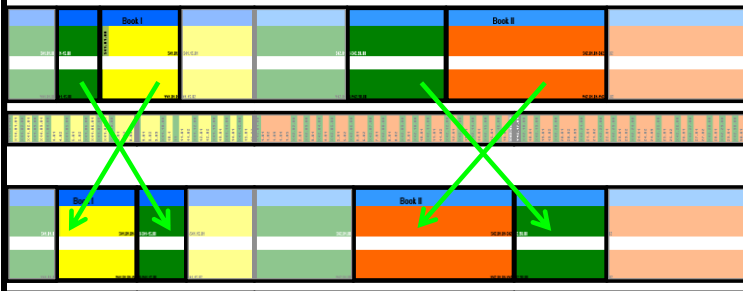
(Slide 45)

To allow us to create segments with comparable contents, we cut each segment of Corpus and Laud in half.
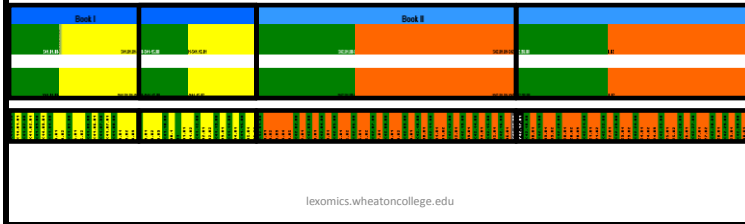

(Slide 46)

We then switch the positions of the sections so that each segment contains *both* the capitulae and the text to which they refer.
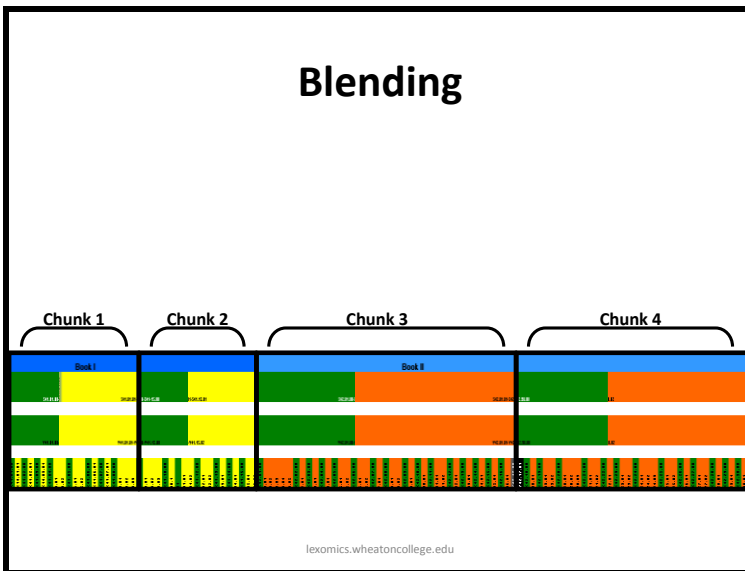
(Slide 47)

It is as if we have taken the table of contents chapter listings and used them for chapter titles in the text (which is what we see in the Junius text).

(Slide 48)

Because the capitulae run in order, when we put the first half of the Corpus and Laud capitulae sections with the first halves of the corresponding texts, the chunks are composed of the same material as the Junius text chunks. We can now compare the dendrograms produced from the blended texts to that produced from the Junius manuscript with some confidence that we are comparing apples to apples and not to oranges.
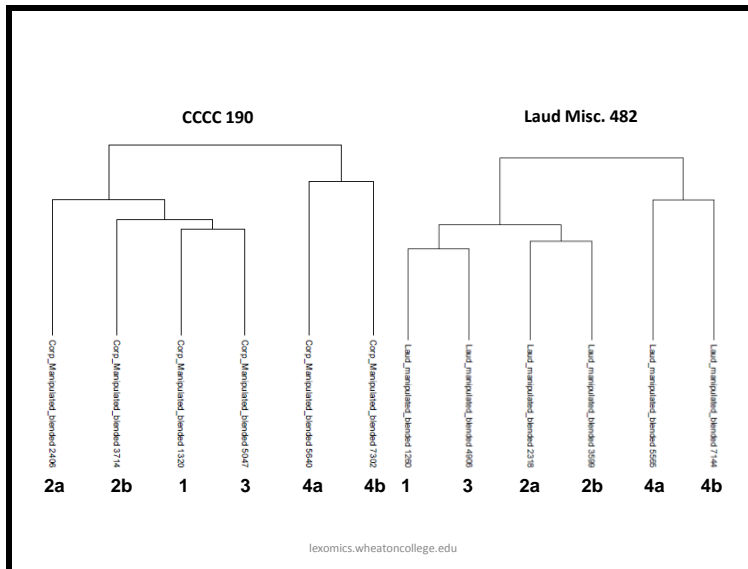
(Slide 49)

Of the three manuscripts, Corpus and Laud have most similar dendrogram geometry, and in the highest level of the clade structure they are the same. Here the chunks are named by their relationship to Books of the *Old English Penitential*. Book One and Three are completely in their own chunks, while Books Two and Four, because they are larger books, are each cut into two chunks labeled "a" and "b".
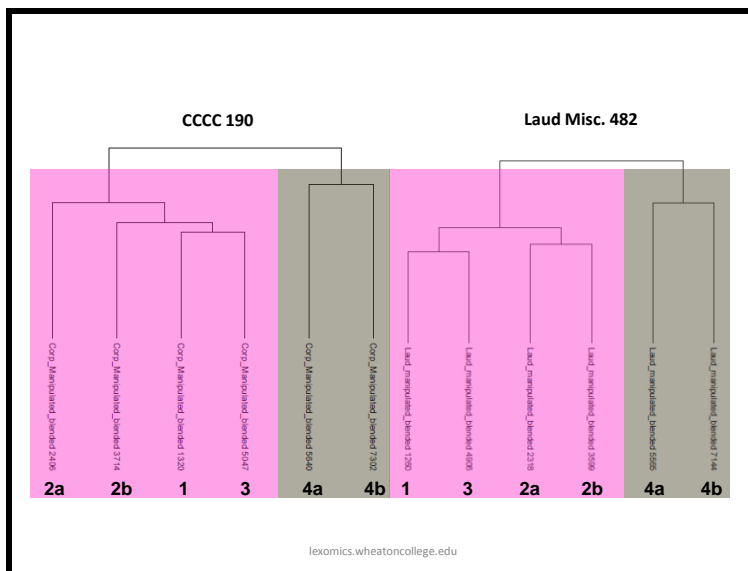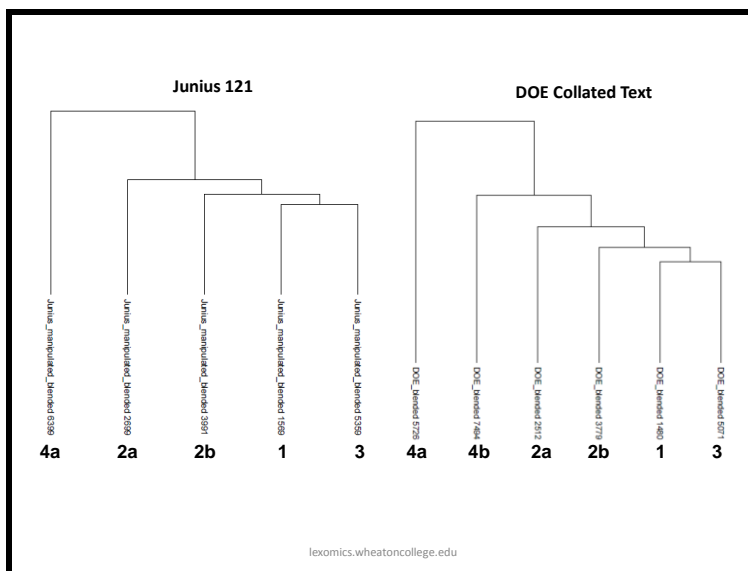

(Slide 50)

Chunks 1 through 2b cluster in one clade with chunks 4a and 4b in the other. The only difference in the two dendrograms is within the large cades highlighted here in pink. In both texts chunk 1 clusters with chunk 3, but in Laud chunks 2a and 2b also cluster together, while in Corpus we see a stepwise geometry with 2a and 2b slightly separate. This is the same difference that we saw between the *Daniel* and *Azarias* dendrograms created from the diplomatic and critical editions. Because the vertical distances between the branches are so short between the inner clades, the geometry may be affected by very small variations in the underlying text.

Junius 121

4a 2a 2b 1 3

lexomics.wheatoncollege.edu
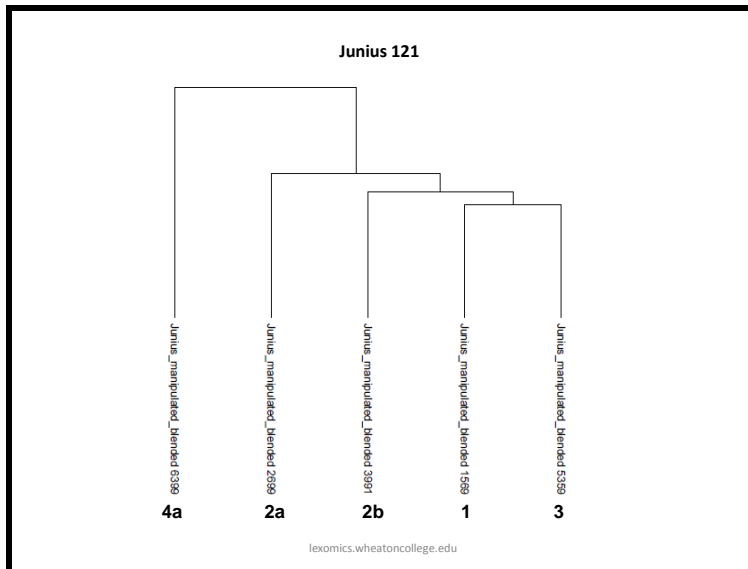
(Slide 51)

When we look at the Junius dendrogram, we see what *appears* to be a different geometry, but when we look more closely, we find that it is in fact basically the same as Corpus and Laud. Note that in Junius chunks 1-3 cluster together in one clade, and that within that clade chunk 1 is most like chunk 3. We then see chunk 4a in its own clade, and this single branch makes the dendrogram look different from Corpus and Laud. But remember that Junius has less text than do the other two manuscripts; half of Book Four is missing. That missing section corresponds with chunk 4b of Laud and Corpus. So the Junius dendrogram has in fact exactly the same geometry as Corpus and nearly the same geometry as Laud.

Junius 121   **DOE Collated Text**

4a 2a 2b 1 3   4a 4b 2a 2b 1 3

lexomics.wheatoncollege.edu

(Slide 52)

It also has the same geometry as the collated edition used by the Dictionary of Old English.
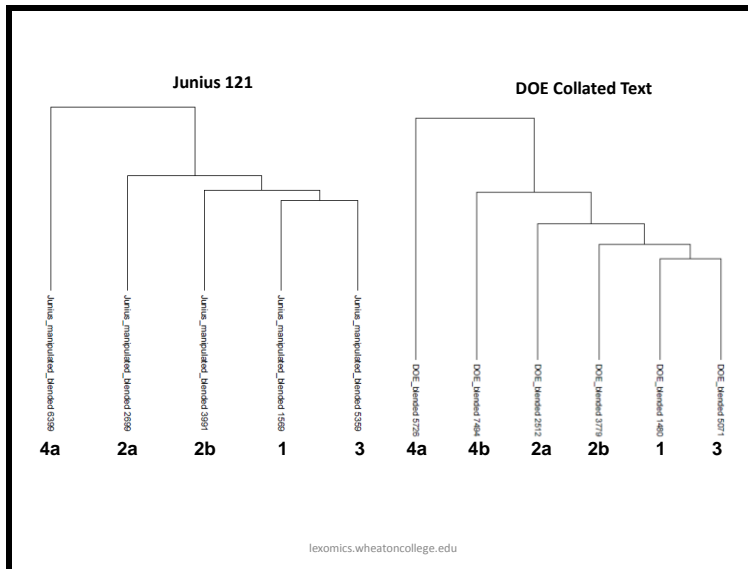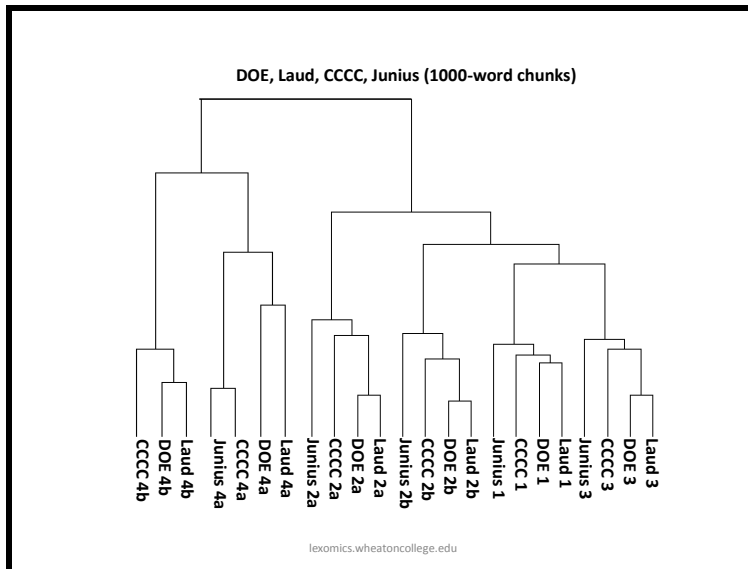
(Slide 53)

The only difference in the two dendrograms is the absence of chunk 4b in the Junius dendrogram. In the Dictionary of Old English dendrogram, this chunk joins the main body of the text and remains separate from chunk 4a, which is the most distinct chunk in both dendrograms. This geometry is consistent with the editorial practice of the critical edition, as Raith used material from Junius to fill in gaps in Corpus and Laud. Raith's combined text therefore makes chunk four a somewhat different from what it is in either Laud or Corpus (where chunk 4a is more similar to chunk 4b).
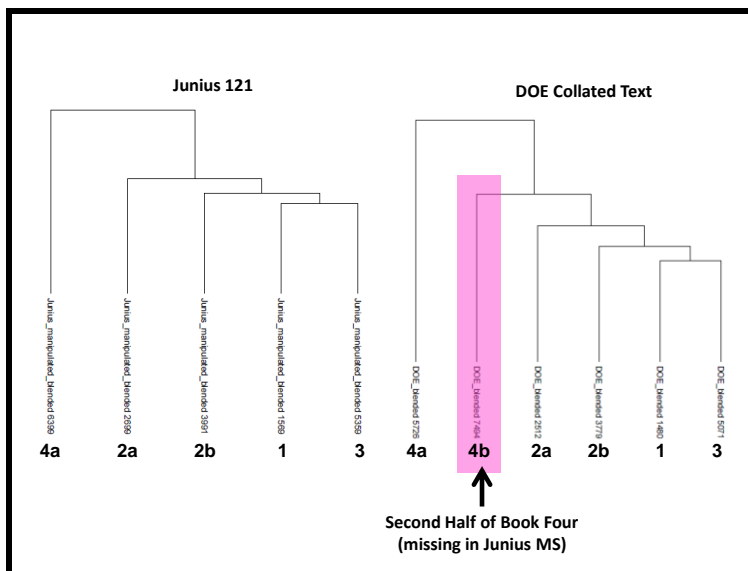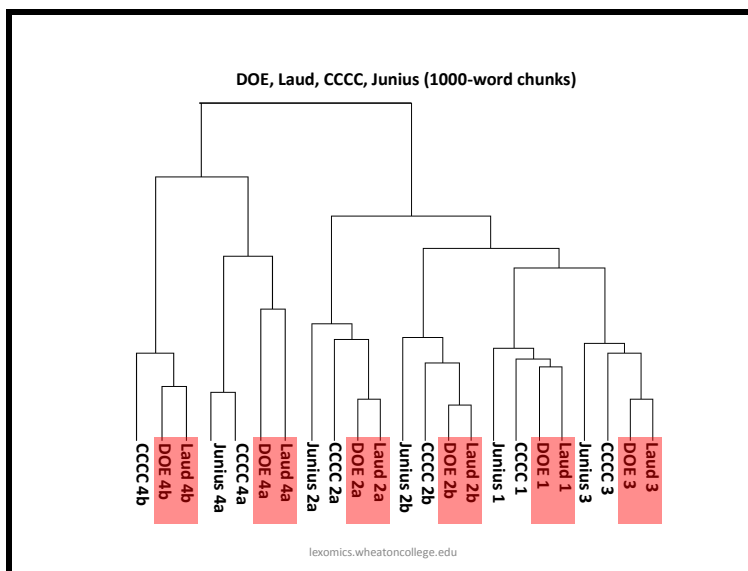

(Slide 54)

To compare the Dictionary of Old English critical edition with the various manuscript editions, we created a dendrogram that includes all four texts.

DOE, Laud, CCCC, Junius (1000-word chunks)

(Slide 55)

What we first see is that chunks of the Dictionary of Old English critical edition stick most closely to the corresponding chunks of the Laud manuscript. This grouping occurs for all chunks, suggesting that the DOE edition follows the vocabulary of the Laud manuscript more closely than it does the other manuscripts.

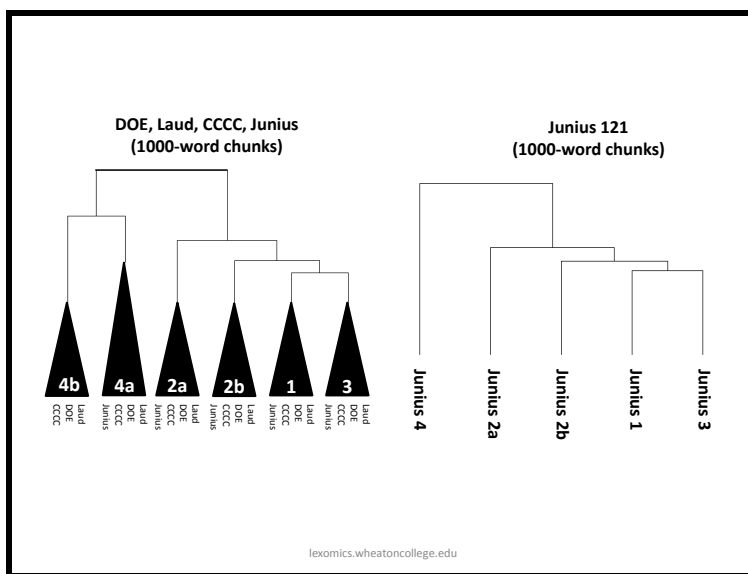DOE, Laud, CCCC, Junius (1000-word chunks)

(Slide 56)

If we simplify the terminal leaves of the dendrogram, we can more easily see the relationships between the texts and the critical edition.
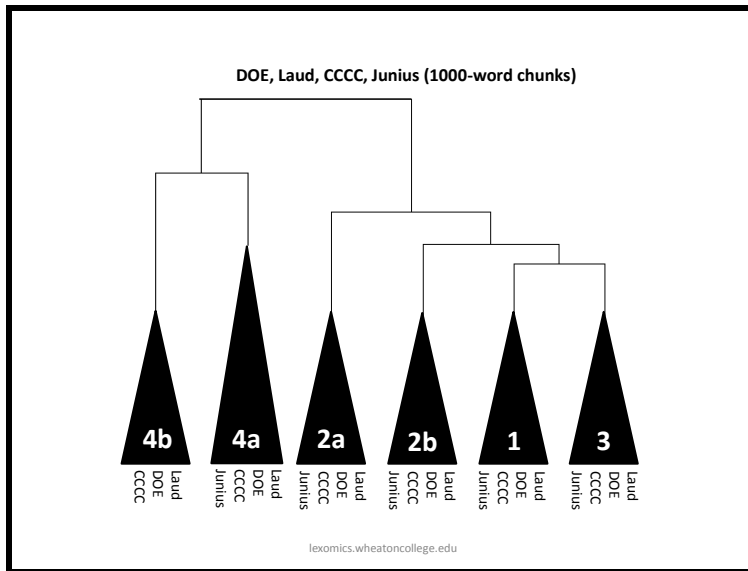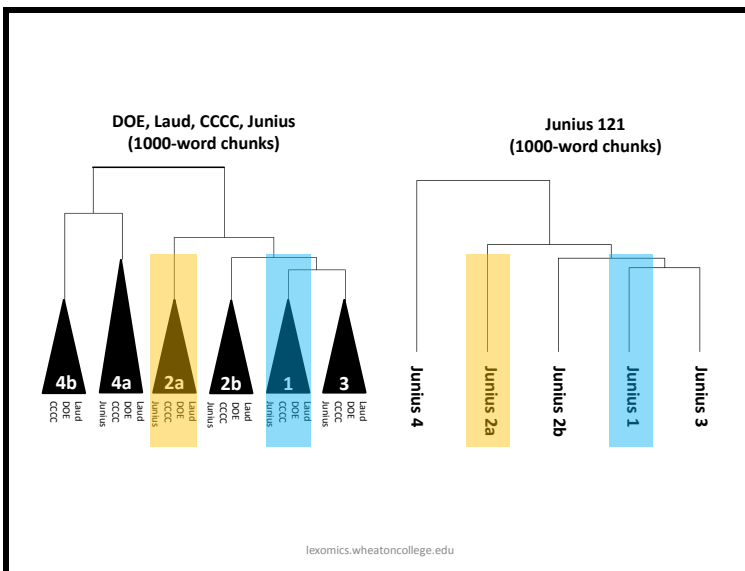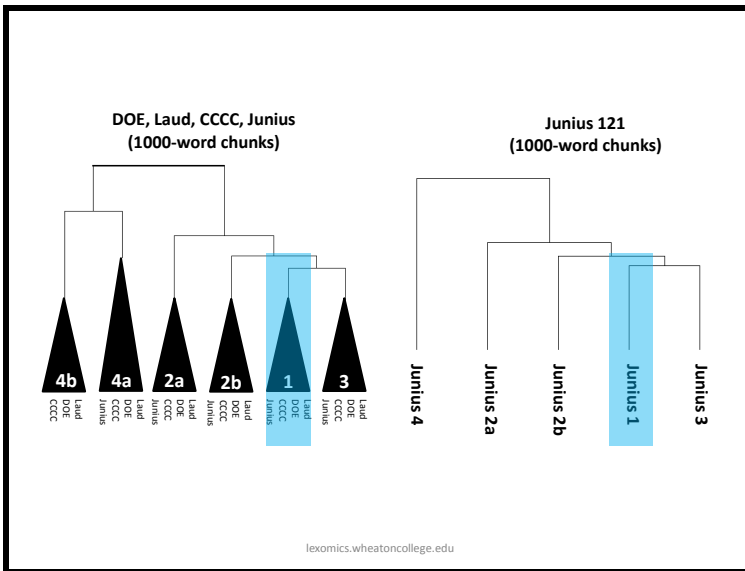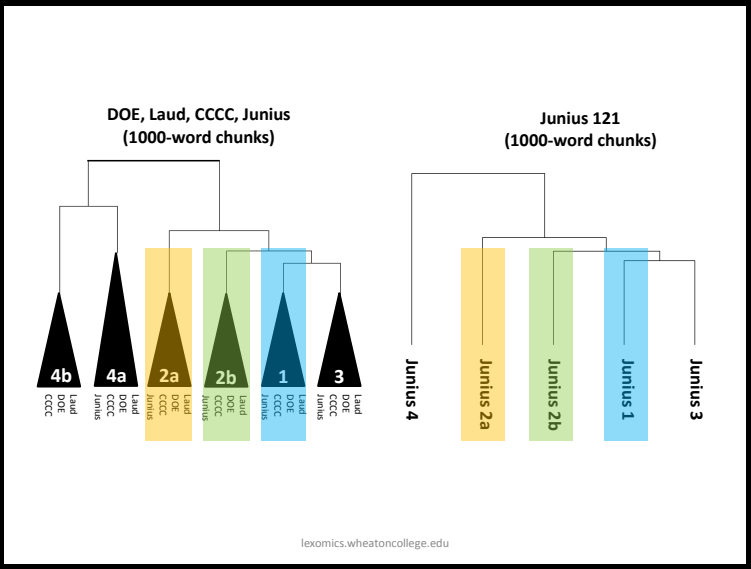
(Slide 57)

Note that the combined dendrogram has the same high-level clade structure as the Junius dendrogram.
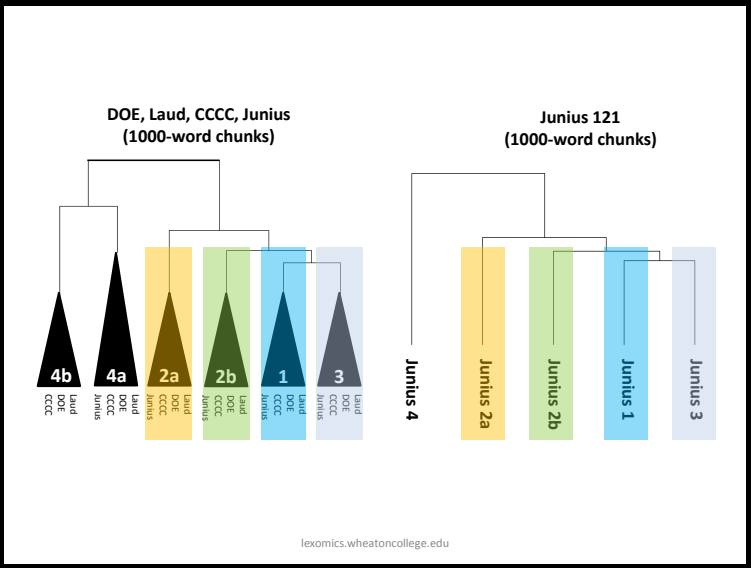

(Slide 58)
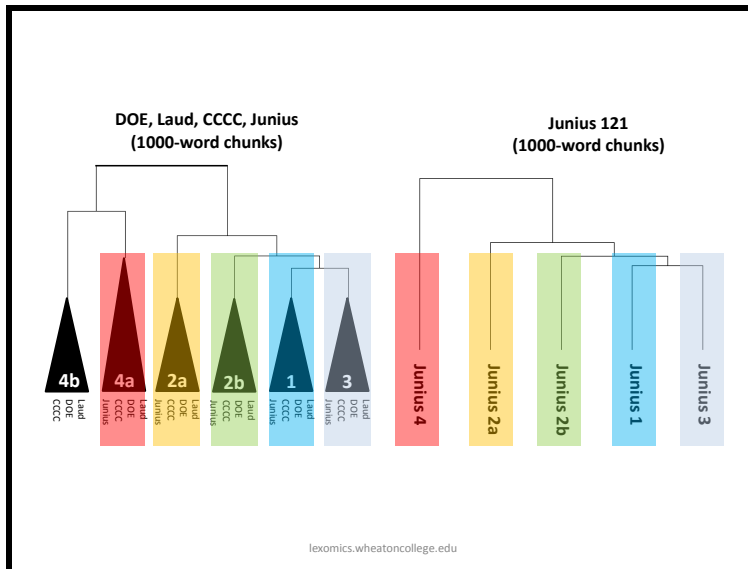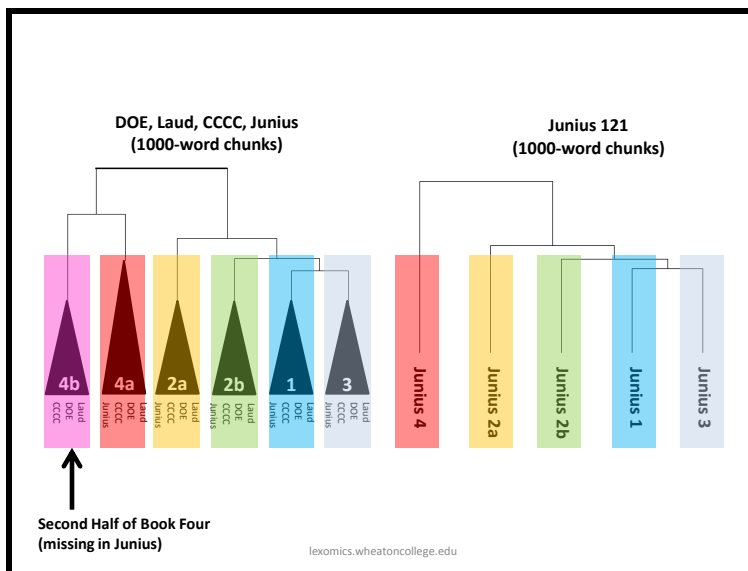
(Slide 59)

(Slide 60)

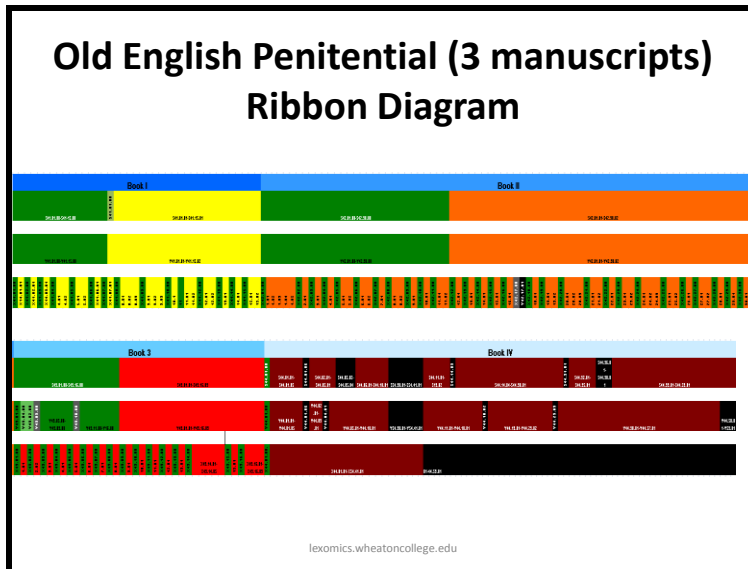(Slide 61)

(Slide 62)

(Slide 63)

Chunk 4b is the only chunk whose placement does not agree with the dendrogram produced from the Junius text.
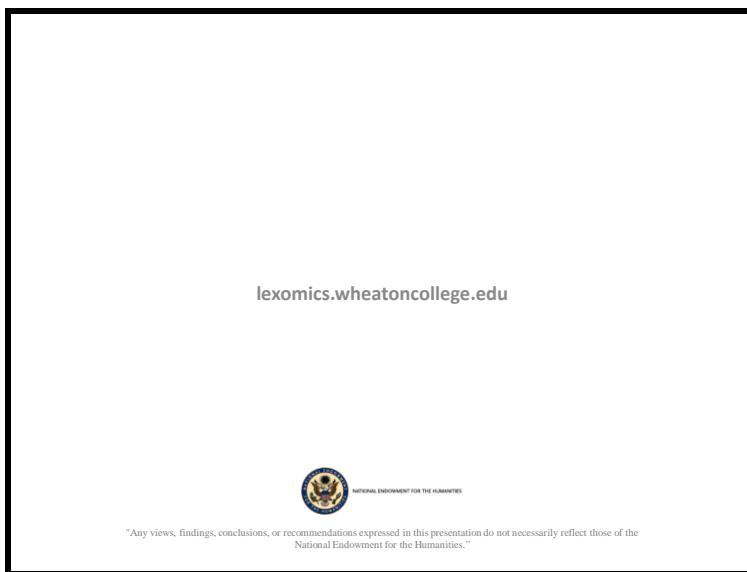

(Slide 64)

4b, remember, is the chunk that contains the material from Book Four of the *Old English Pentitential*, which is missing from the Junius manuscript.

Old English Penitential (3 manuscripts)
Ribbon Diagram

lexomics.wheatoncollege.edu

(Slide 65)

Looking back at the ribbon, you can see this section in black.

lexomics.wheatoncollege.edu

NATIONAL ENDOWMENT FOR THE HUMANITIES

"Any views, findings, conclusions, or recommendations expressed in this presentation do not necessarily reflect those of the National Endowment for the Humanities."

(Side 66)

From all of these experiments we can draw several conclusions. First, at the higher levels of the clade structure, there is no significant difference between the dendrograms produced from diplomatic and critical editions. We can therefore use either and still get results that agree with the controls.

There are small differences in structure between diplomatic and critical editions at lower levels of the clade structure. In every case that we have examined, the diplomatic edition seems less reliable in Lexomic terms than the critical edition. We believe this to be the case because the diplomatic editions repeat scribal errors, introducing error into the word counts upon which Lexomic analysis relies.

Even with collated editions, like that of the penitentials, we find that the critical editions and the diplomatic editions are not significantly different once we take into account how the editor put together the edition. If

we believe that the ultimate exemplar of the *Old English Penitential* was something that included the first three books translated from Haltigar plus the fourth book taken from the Old English *Scriftboc*, we see that the critical edition accurately reflects this hypothesized exemplar, and the dendrograms taken from this edition also demonstrate the same basic clade structure as the manuscripts.

We still must pay close attention to editorial practice and make sure we are comparing like to like when we produce a dendrogram from a critical edition, but we can be reasonably confident that analysis based on the critical editions in the Dictionary of Old English corpus is consistent with what could be produced from diplomatic electronic editions. The extra labor of producing diplomatic electronic editions, then, seems not worth the payoff. The critical editions are sufficient for the purposes of Lexomic analysis.

By: Michael Drout and Leah Smith

lexomics.wheatoncollege.edu