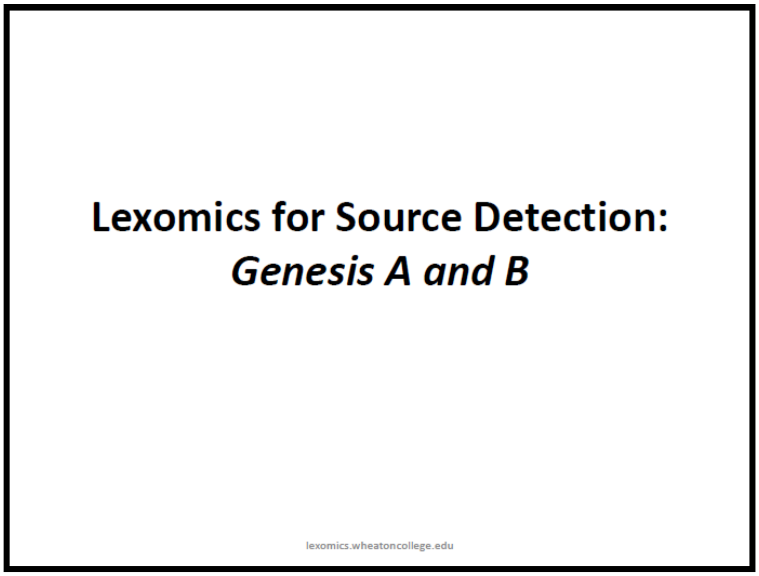
## Lexomics for Source Detection

### Source Detection in Anglo-Saxon Poetry



Lexomics for Source Detection:
*Genesis A and B*

lexomics.wheatoncollege.edu

(Slide 1)

Lexomics can be used to identify vocabulary variation within texts. As we discussed in "The Story of *Daniel*," in that Anglo-Saxon poem the vocabulary differences reflected in the dendrogram match philologists' theories about the poet's use of both the Latin Bible and Latin canticles as sources. We were therefore able to use Lexomic methods to detect the influence of different sources on different parts of the text.

### *Genesis A* and *B*

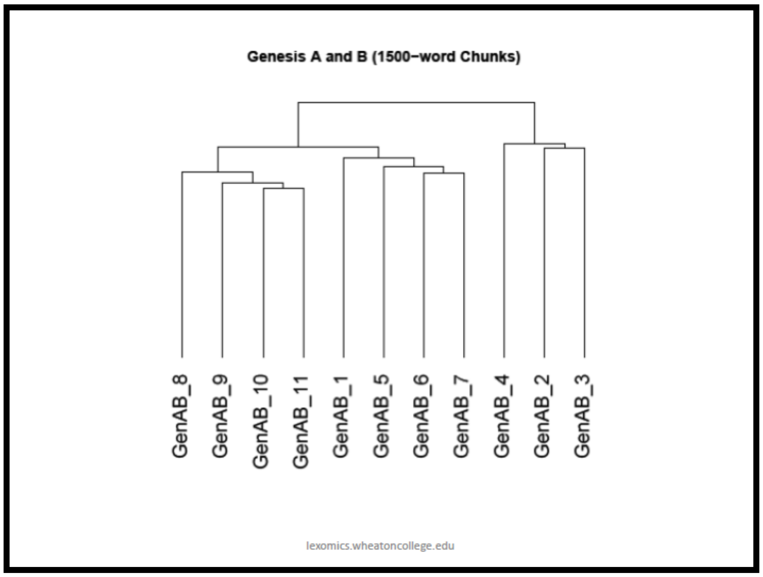
*Genesis* from the Junius Manuscript, Anglo-Saxon

(Slide 2)

The Anglo-Saxon poem Genesis appears to be an Old English poetic paraphrase of much of the biblical book of Genesis from the Latin Bible. But in 1875 Eduard Sievers used philological evidence to argue that lines 235-851 were a translation of an Old Saxon original. Nineteen years later a manuscript was found in the Vatican Library that proved Sievers' deduction was correct.

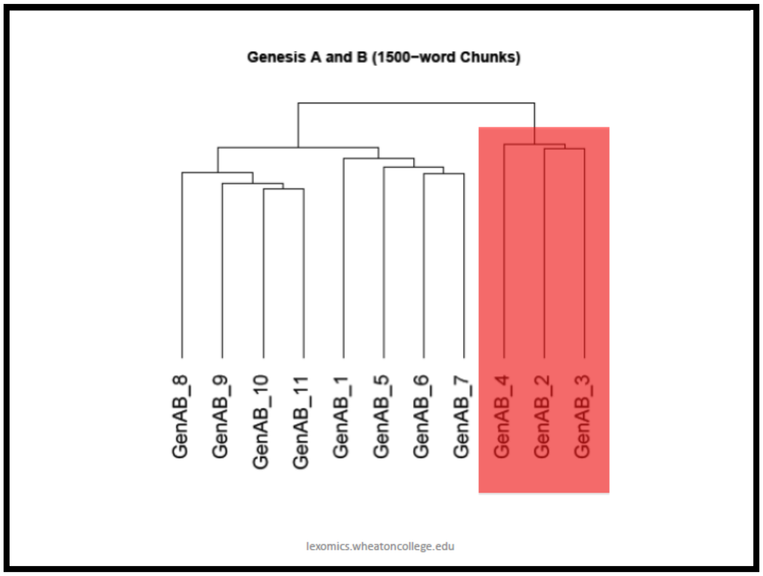
Lines 235-851 of the poem are now called *Genesis B*, while the lines directly translated from the biblical Latin are called *Genesis A.*

Mostly subtle distinctions in spelling and meter, the differences between *Genesis A* and *Genesis B* are not immediately obvious to someone without philological training. But in the dendrogram we created using Lexomic methods the two sections of the poem are clearly separated.



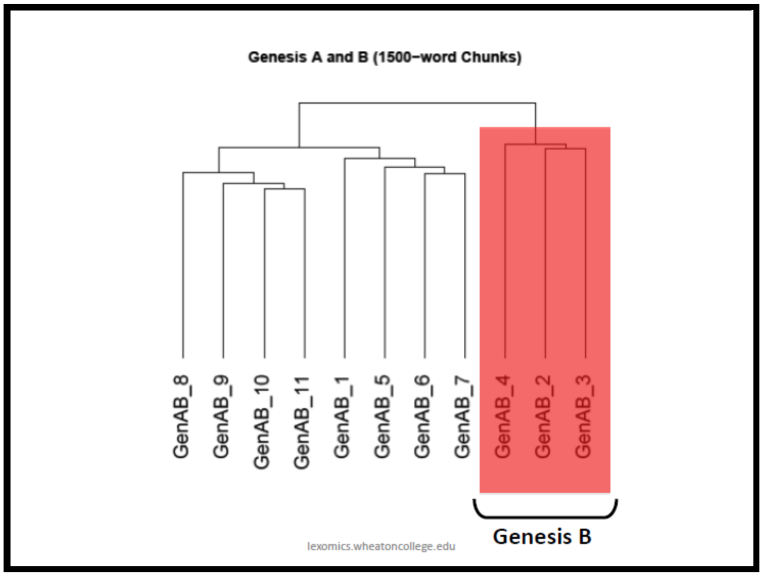
(Slide 3)

We cut the poem into 1500-word segments, and then produced this dendrogram.



(Slide 4)

Chunks two, three, and four form a separate clade, very distinct from the rest of the poem.

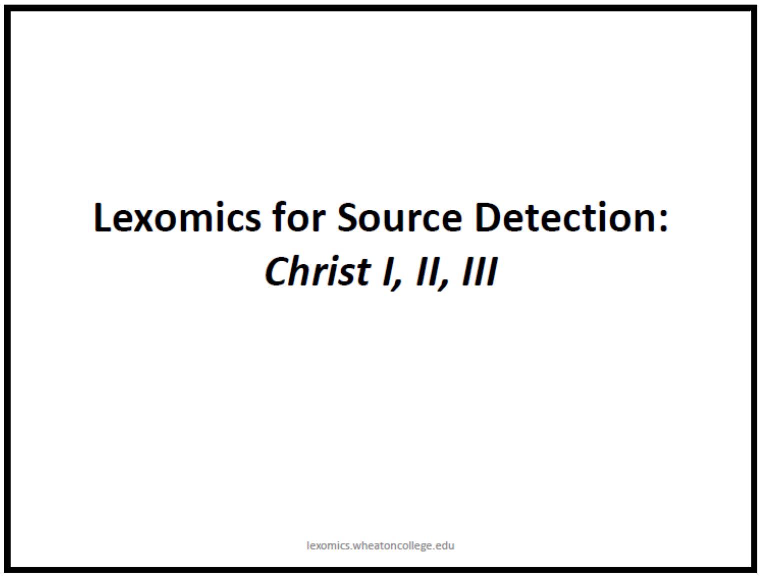


(Slide 5)

These three chunks contain the Genesis B text.

The dendrograms of *Daniel* and *Genesis* show that the methods can detect portions of an Old English text that have different sources from the body of the main text.

***Christ I, II,* and *III***

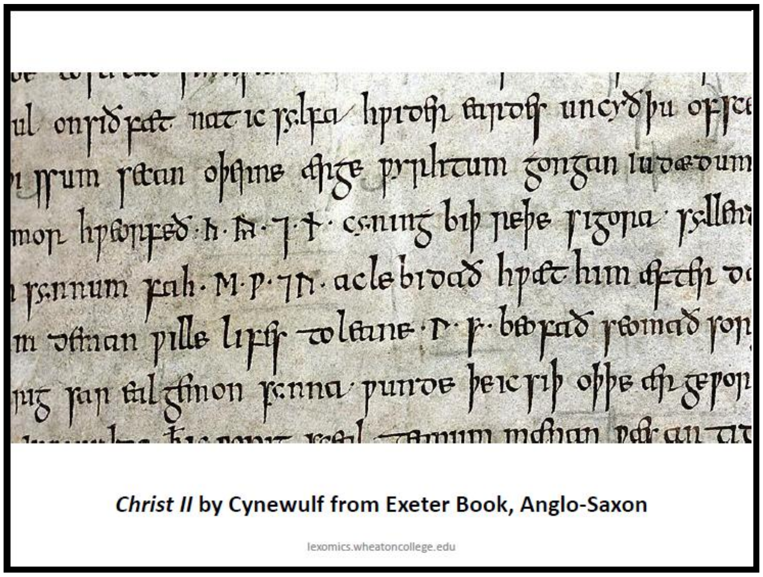


Lexomics for Source Detection:
*Christ I, II, III*

lexomics.wheatoncollege.edu

(Slide 6)

This conclusion is further supported by Lexomic analysis of *Christ I, II,* and *III*, the first three poems in the tenth-century manuscript known as the Exeter Book.

(Slide 7) 

***Christ II* by Cynewulf from Exeter Book, Anglo-Saxon**

lexomics.wheatoncollege.edu

These texts were once thought to be one long poem about Christ, but for the past century scholars have agreed that they are three separate poems brought together by the compiler of the manuscript. The subject of *Christ I* is the Advent, *Christ II*, the Ascension, and *Christ III* the Last Judgment. We originally wanted to use Lexomic analysis to test this hypothesis and see if the Christ poems would separate out into their own clades.

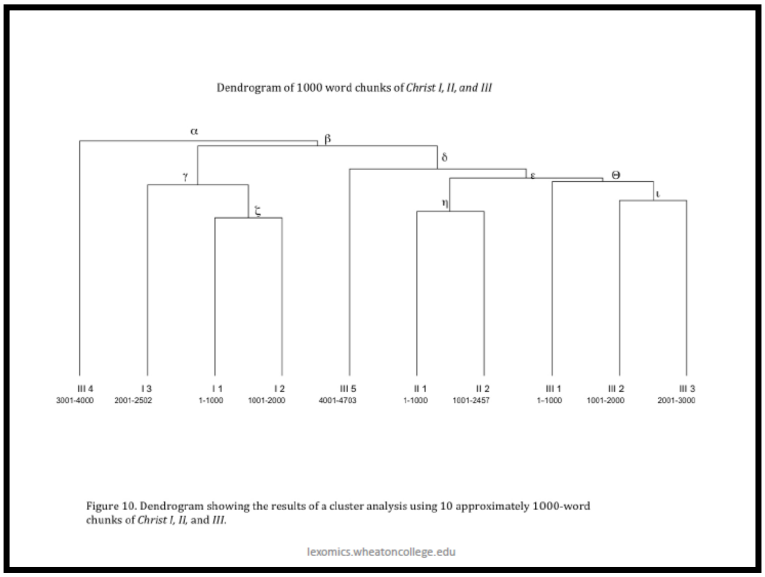
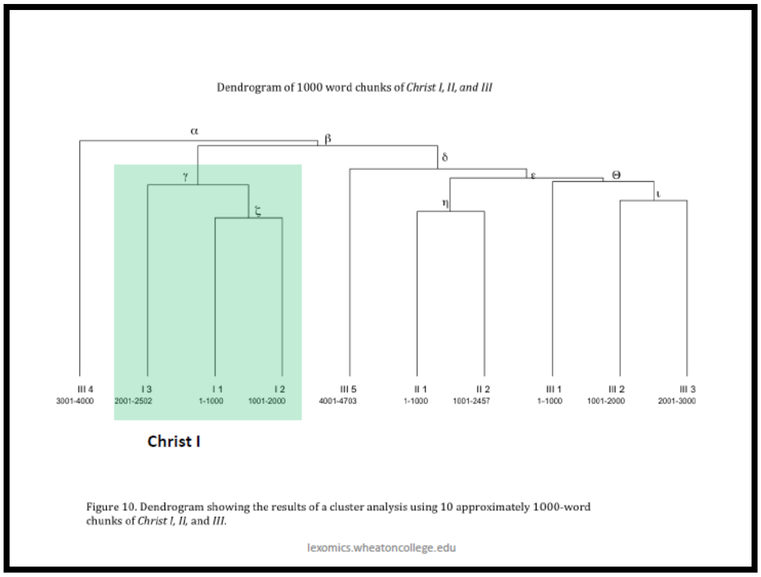
(Slide 8) 

Figure 10. Dendrogram showing the results of a cluster analysis using 10 approximately 1000-word chunks of *Christ I, II,* and *III*.

lexomics.wheatoncollege.edu

We cut all three poems into segments of approximately 1000 words and then produced this dendrogram.

(Slide 9)



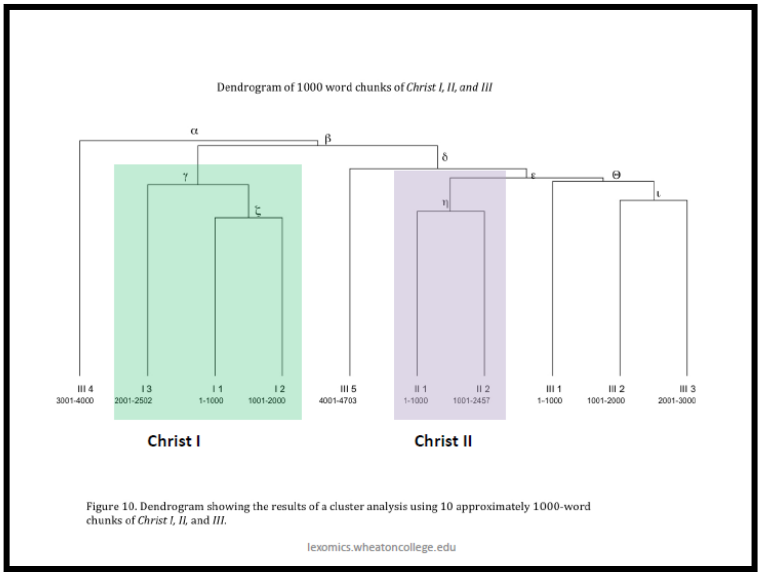
Dendrogram of 1000 word chunks of *Christ I, II, and III*

Figure 10. Dendrogram showing the results of a cluster analysis using 10 approximately 1000-word chunks of *Christ I, II,* and *III.*

lexomics.wheatoncollege.edu

Note first that Lexomic methods correctly separate the three poems, producing one clade that is made up only of chunks of *Christ I*,

(Slide 10)



Dendrogram of 1000 word chunks of *Christ I, II, and III*

Figure 10. Dendrogram showing the results of a cluster analysis using 10 approximately 1000-word chunks of *Christ I, II,* and *III.*

lexomics.wheatoncollege.edu

…another that is only *Christ II*,

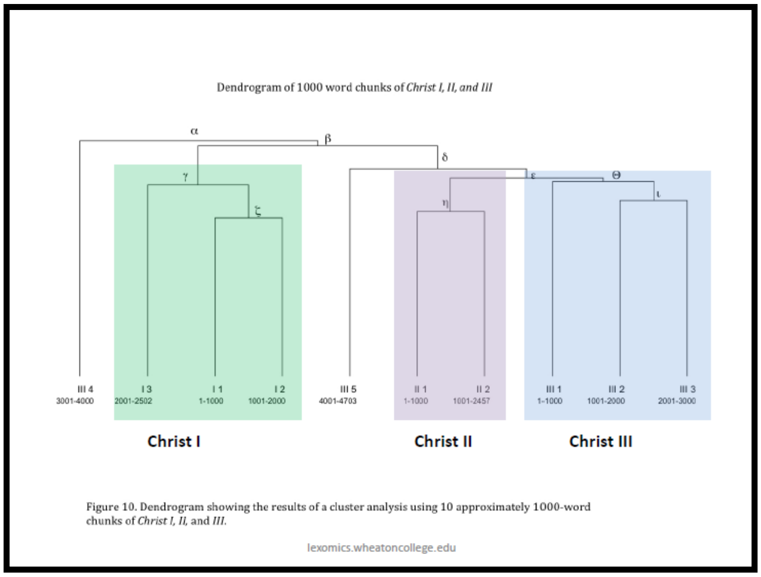

Dendrogram of 1000 word chunks of *Christ I, II, and III*

Figure 10. Dendrogram showing the results of a cluster analysis using 10 approximately 1000-word chunks of *Christ I, II,* and *III*.

lexomics.wheatoncollege.edu

(Slide 11)

…and a third that is only *Christ III*. These three groupings are indicated by the different colored squares superimposed on the dendrogram.

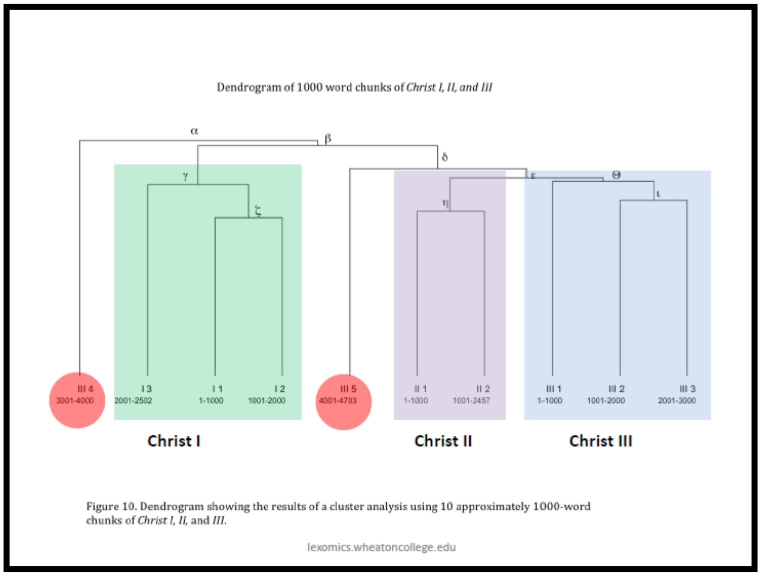
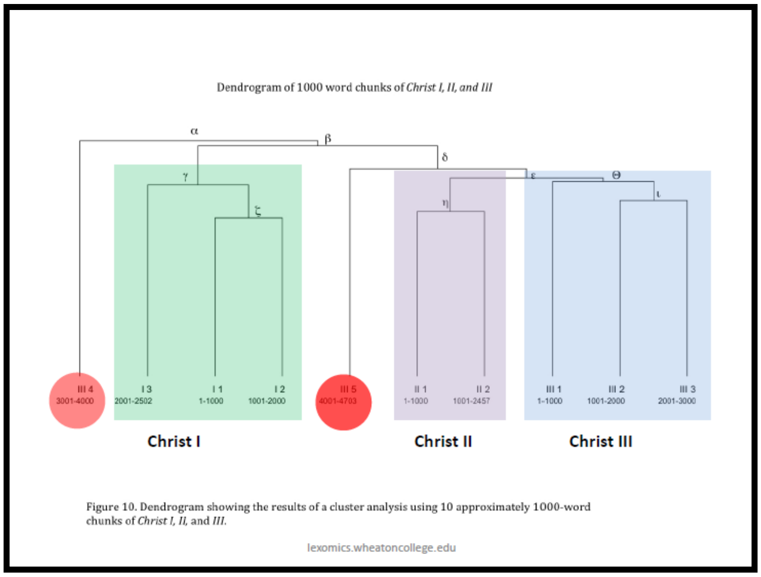

Dendrogram of 1000 word chunks of *Christ I, II, and III*

Figure 10. Dendrogram showing the results of a cluster analysis using 10 approximately 1000-word chunks of *Christ I, II,* and *III*.

lexomics.wheatoncollege.edu

(Slide 12)

However, two segments of *Christ III*, chunks four and five, are separate from the rest of the *Christ III* grouping.



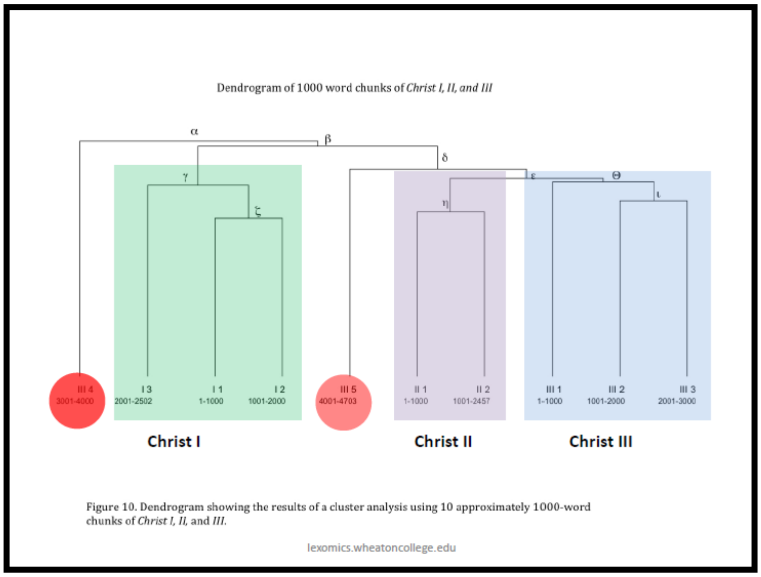
Dendrogram of 1000 word chunks of *Christ I, II, and III*

Figure 10. Dendrogram showing the results of a cluster analysis using 10 approximately 1000-word chunks of *Christ I, II,* and *III*.

lexomics.wheatoncollege.edu

(Slide 13)

Chunk five is closer to the rest of the *Christ III* grouping than chunk four,



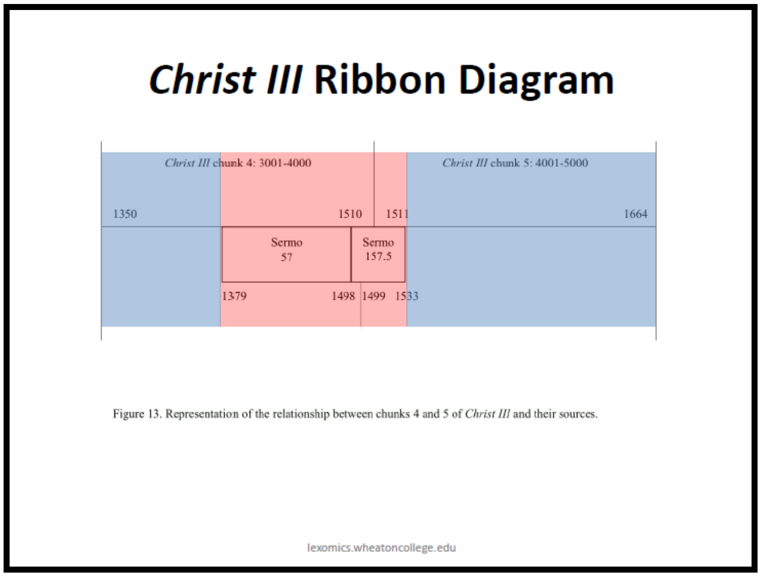
Dendrogram of 1000 word chunks of *Christ I, II, and III*

Figure 10. Dendrogram showing the results of a cluster analysis using 10 approximately 1000-word chunks of *Christ I, II,* and *III*.

lexomics.wheatoncollege.edu

(Slide 14)

…which is isolated in its own clade, separate from every other chunk in all three Christ poems. Why do these two chunks have different vocabulary distributions than the rest of the poems?



## *Christ III* Ribbon Diagram

*Christ III* chunk 4: 3001-4000
*Christ III* chunk 5: 4001-5000
1350
1510
1511
1664
Sermo 57
Sermo 157.5
1379
1498
1499
1533

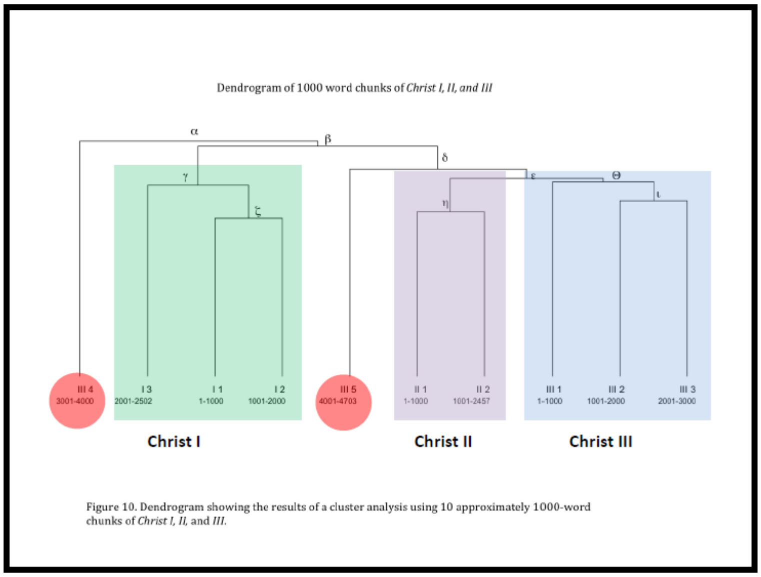
Figure 13. Representation of the relationship between chunks 4 and 5 of *Christ III* and their sources.

lexomics.wheatoncollege.edu

(Slide 15)

The simple answer is that these two segments are influenced by outside sources, just like the segments of *Daniel* and *Genesis* are. This ribbon diagram of *Christ III* shows that – as Albert S. Cook first noted over a century ago – most of chunk 4 of *Christ III* is an adaptation of a Latin text, Sermo 57 of Caesarius of Arles. Chunk 5 contains some material from a different Latin text by Caesarius, Sermo 157.5.

We conclude that chunk 4 moves further away from the rest of *Christ III* than does chunk 5 because it includes more material from these external sources: 77.5% of chunk four has its source in Caesarius, while only 10% of chunk 5 comes from that Latin source.



Dendrogram of 1000 word chunks of *Christ I, II, and III*

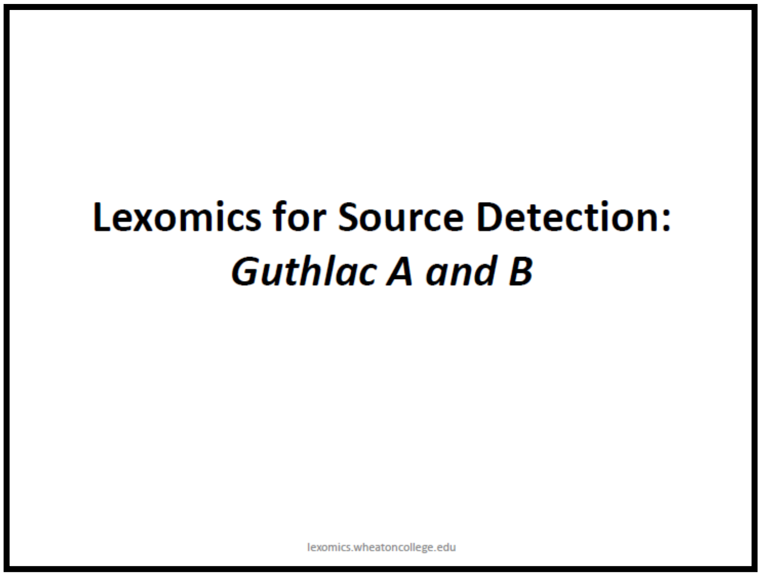
Figure 10. Dendrogram showing the results of a cluster analysis using 10 approximately 1000-word chunks of *Christ I, II,* and *III.*

(Slide 16)

This analysis of the *Christ* poems is further evidence that Lexomic methods can be used to detect sources of texts.

Analysis of the Old English *Guthlac* poems allows us to refine further the techniques.

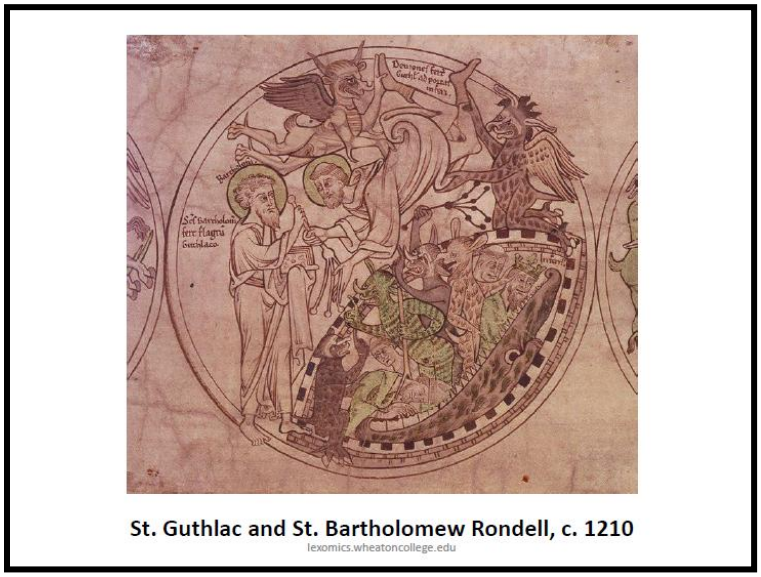
***Guthlac A* and *B***

(Slide 17) 

Lexomics for Source Detection:
*Guthlac A and B*

lexomics.wheatoncollege.edu

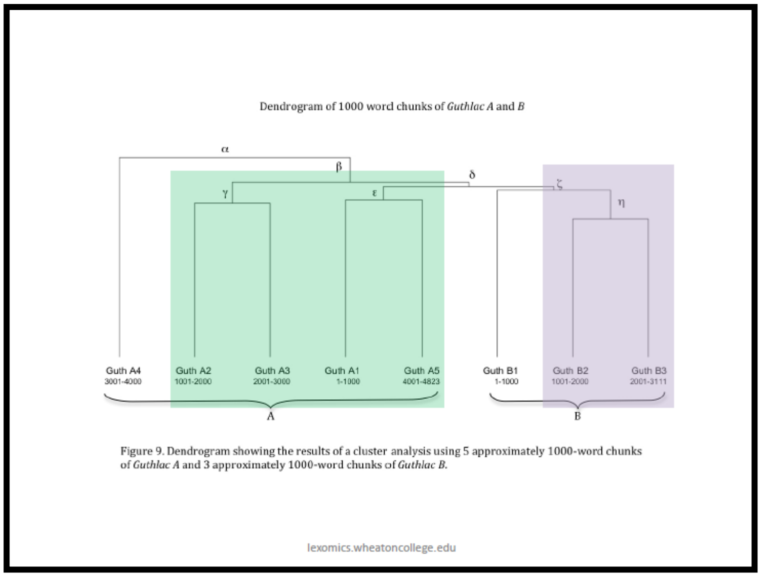
The Old English poems *Guthlac A* and *B* immediately follow the *Christ* poems in the Exeter Book.

(Slide 18) 

St. Guthlac and St. Bartholomew Rondell, c. 1210

lexomics.wheatoncollege.edu

Both tell the story of Guthlac, a seventh-century Anglo-Saxon saint from the Fenlands. The earliest source we have for the life of Guthlac is the *Vita Sancti Guthlaci* by Felix, an eighth century Latin text. This text was translated into Old English prose, from which a short excerpt was taken to produce *Vercelli Homily 23*. The Latin *Vita* is also somehow related to *Guthlac A* and is the source of most of *Guthlac B*.



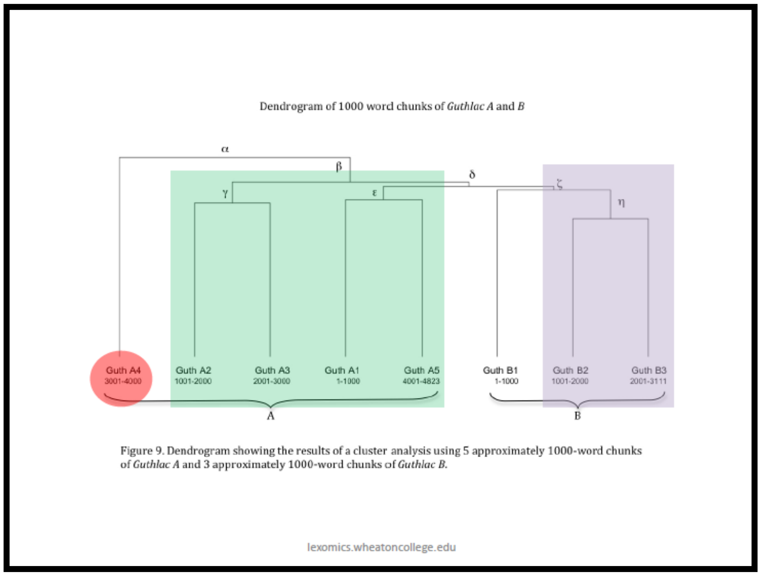
Dendrogram of 1000 word chunks of *Guthlac A* and *B*

Figure 9. Dendrogram showing the results of a cluster analysis using 5 approximately 1000-word chunks of *Guthlac A* and 3 approximately 1000-word chunks of *Guthlac B*.

lexomics.wheatoncollege.edu

(Slide 19)

As this dendrogram shows, Lexomic analysis is able to separate the two *Guthlac* poems from each other: *Guthlac A* and *Guthlac B* are marked with green and purple boxes.

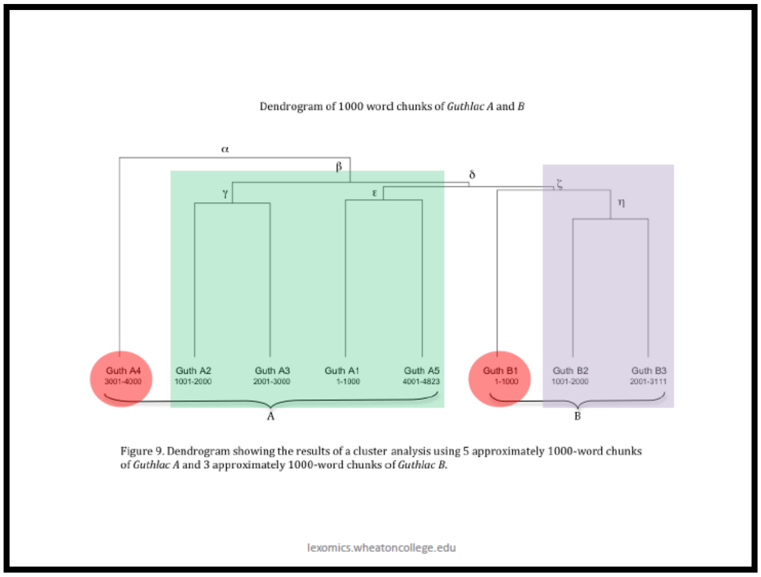


Dendrogram of 1000 word chunks of *Guthlac A* and *B*

Figure 9. Dendrogram showing the results of a cluster analysis using 5 approximately 1000-word chunks of *Guthlac A* and 3 approximately 1000-word chunks of *Guthlac B*.

lexomics.wheatoncollege.edu

(Slide 20)

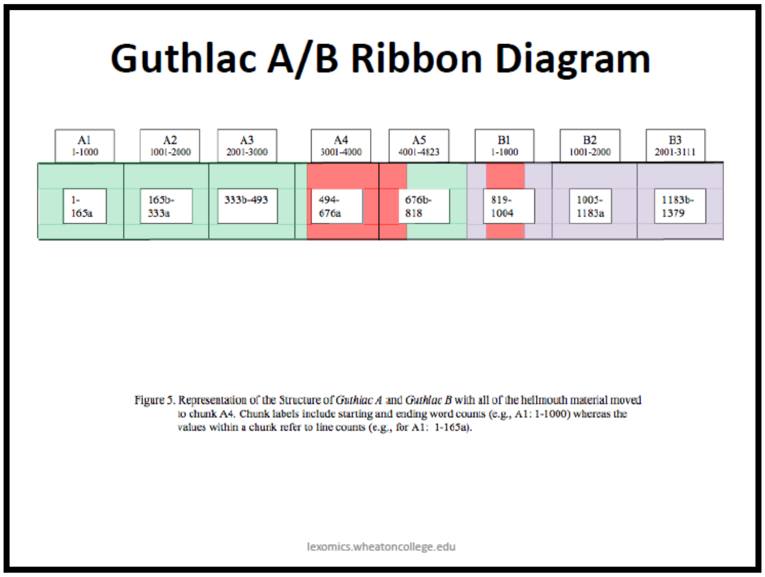
But note how the fourth segment of *Guthlac A* is on its own branch all the way to one side of the dendrogram. *Guthlac A*4 is therefore different in vocabulary from every other segment of either poem. And while *Guthlac B* all does go together in a single clade, one chunk, (Slide 21) 

Dendrogram of 1000 word chunks of *Guthlac A* and *B*

Figure 9. Dendrogram showing the results of a cluster analysis using 5 approximately 1000-word chunks of *Guthlac A* and 3 approximately 1000-word chunks of *Guthlac B*.

lexomics.wheatoncollege.edu

…*Guthlac B*1, is separate from the other two.

(Slide 22) 

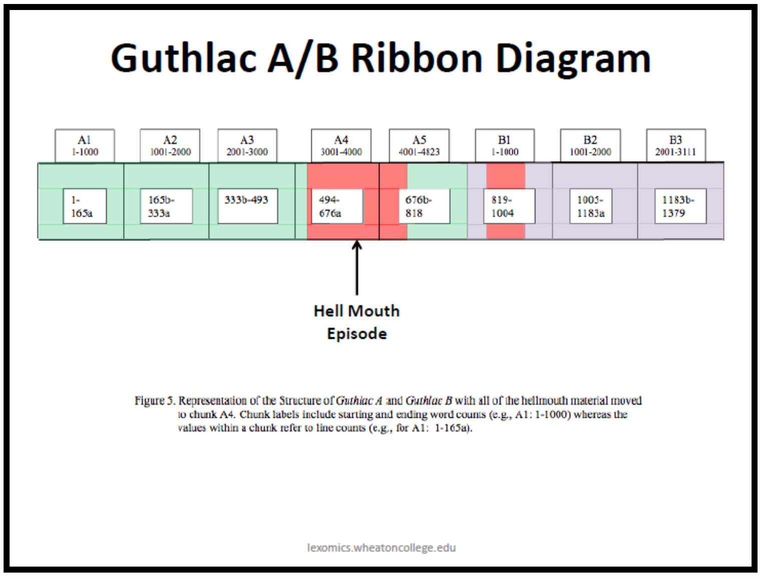
## Guthlac A/B Ribbon Diagram

| A1<br>1-1000 | A2<br>1001-2000 | A3<br>2001-3000 | A4<br>3001-4000 | A5<br>4001-4823 | B1<br>1-1000 | B2<br>1001-2000 | B3<br>2001-3111 |
|---|---|---|---|---|---|---|---|
| 1-165a | 165b-333a | 333b-493 | 494-676a | 676b-818 | 819-1004 | 1005-1183a | 1183b-1379 |

Figure 5. Representation of the Structure of *Guthlac A* and *Guthlac B* with all of the hellmouth material moved to chunk A4. Chunk labels include starting and ending word counts (e.g., A1: 1-1000) whereas the values within a chunk refer to line counts (e.g., for A1: 1-165a).

lexomics.wheatoncollege.edu

If we produce a ribbon diagram of the two poems, we can identify the content within these divergent chunks.



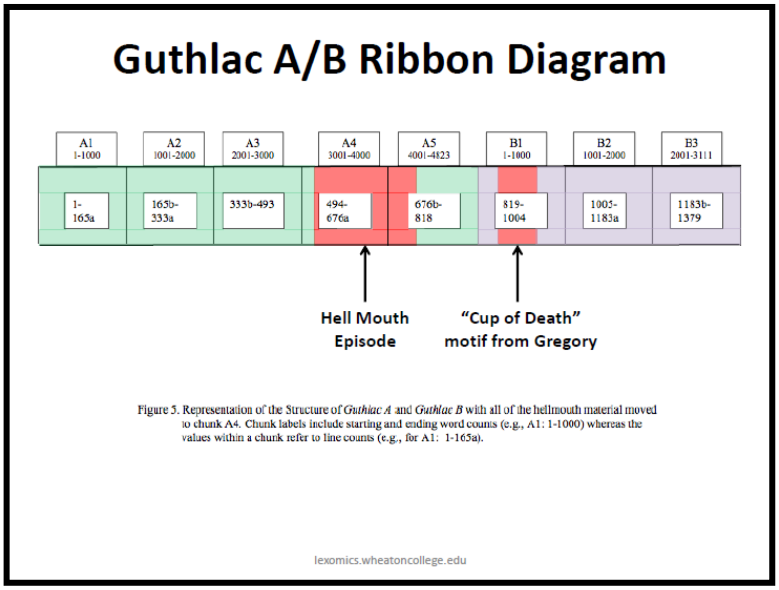
Guthlac A/B Ribbon Diagram

Figure 5. Representation of the Structure of *Guthlac A* and *Guthlac B* with all of the hellmouth material moved to chunk A4. Chunk labels include starting and ending word counts (e.g., A1: 1-1000) whereas the values within a chunk refer to line counts (e.g., for A1: 1-165a).

lexomics.wheatoncollege.edu

(Slide 23)

Chunk A4 contains a famous episode in which demons grab Guthlac, drag him to the mouth of hell, and threaten to throw him in.



Guthlac A/B Ribbon Diagram

Figure 5. Representation of the Structure of *Guthlac A* and *Guthlac B* with all of the hellmouth material moved to chunk A4. Chunk labels include starting and ending word counts (e.g., A1: 1-1000) whereas the values within a chunk refer to line counts (e.g., for A1: 1-165a).

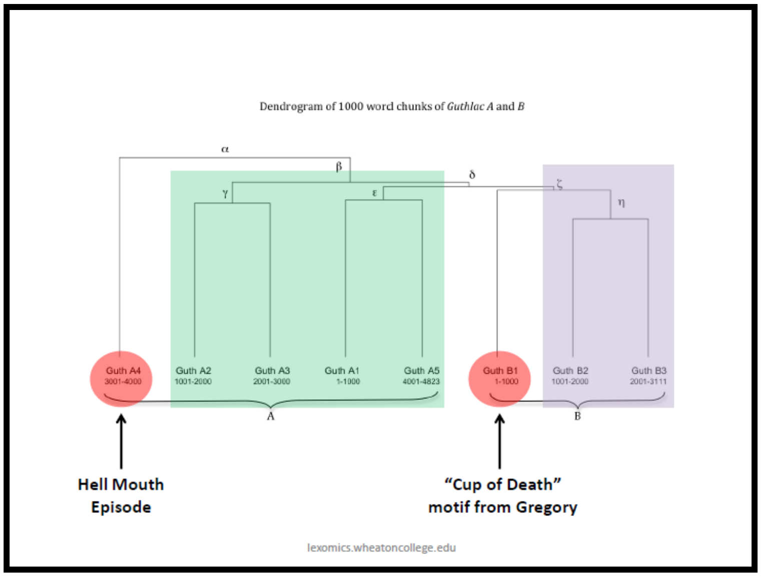
lexomics.wheatoncollege.edu

(Slide 24)

Chunk *Guthlac B*1 contains introductory material that includes a "Cup of Death" motif that circulated widely in the Middle Ages but is not part of Felix's Vita, which is the source for the rest of *Guthlac B.*

The hellmouth episode is exactly the scene that is dramatized in the short prose text we just mentioned, *Vercelli Homily 23*.

(Slide 25)



Hell Mouth Episode

"Cup of Death" motif from Gregory

lexomics.wheatoncollege.edu

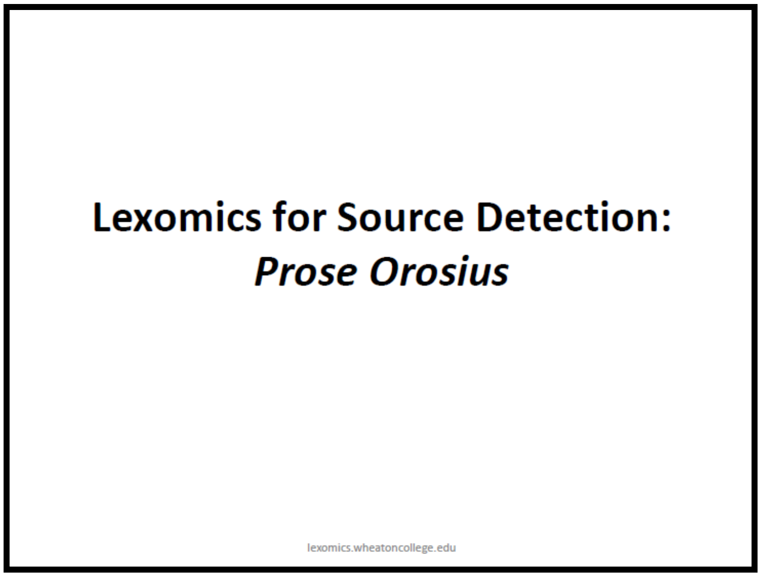
Lexomic analysis of *Daniel, Genesis* and the *Christ* poems suggests that when we see a clade significantly separated from the rest of the dendrogram, that clade has a different source than does the main body of the text. We can conclude, then, that chunk *Guthlac A*4 has such a source, and since we have a text—*Vercelli 23*—which matches that material, we can further conclude that something similar in content to *Vercelli 23* came first and influenced *Guthlac A*4.

Traditional analysis further supports this conclusion. If we read the poem carefully, we note that the poet, just before starting the hellmouth episode, mentions that he learned what he is telling "from books." Rather than taking this as a commonplace or a reference to the Bible, we should instead read the line as an open acknowledgment that the poet has a particular source for what is coming: a stand-alone version of the hellmouth episode like that which is preserved in *Vercelli 23*.

The poet of *Guthlac B* also refers to a book, right at the end of the passage that contains the cup of death motif and just before he begins basing his poem on Felix's Vita, saying "books tell us how Guthlac, through God's will, became blessed in England." Note how this segment is separate from the rest of the poem, indicating that it has a different source.

We can conclude that Lexomic methods give us a new way of investigating the structure and sources of Anglo-Saxon poetic texts. The geometry of a dendrogram can be used to indicate where different sources produce different vocabulary distributions. Combined with a ribbon diagram, a dendrogram can give us hints about where to look for information that might otherwise not have seemed significant.

**Source Detection in Anglo-Saxon Prose**



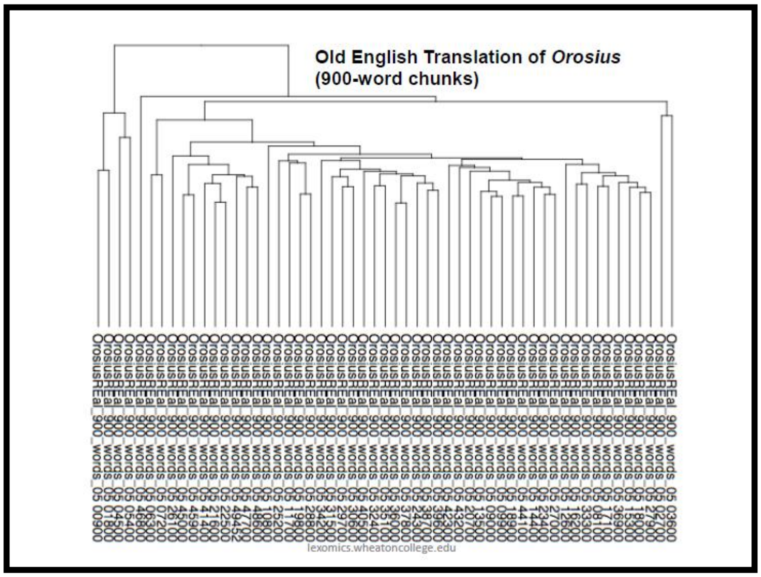
**Lexomics for Source Detection:**
***Prose Orosius***

lexomics.wheatoncollege.edu

(Slide 26)

But Old English *poetry* is different from Old English *prose*: the poetry is more compressed, it uses a particular and limited vocabulary, and it follows rules of meter and alliteration. We could not just assume that the same methods that work on poetry would work on prose. So we needed to test the methods with prose texts whose structures and sources had been convincingly determined using traditional methods.

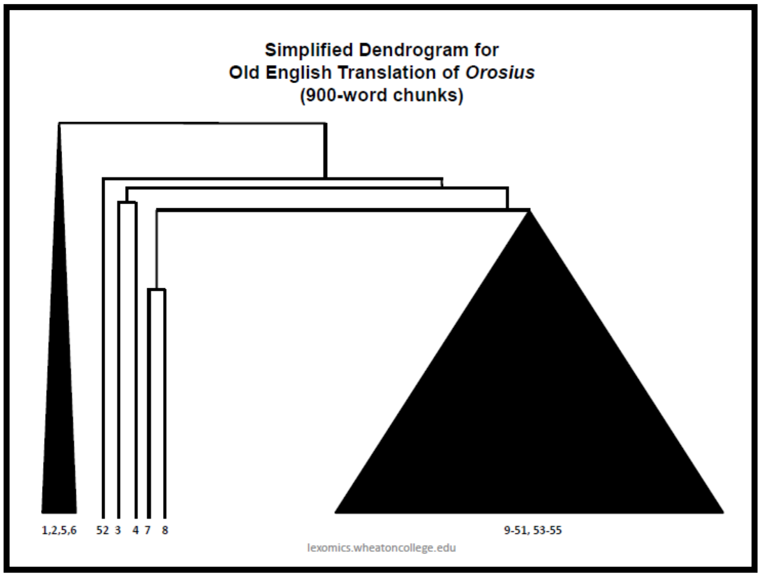
Let us look at the Old English translation of Orosius' *History Against the Pagans*. Although no longer credited to King Alfred himself, the translation was made some time during the Alfredian period. The Old English *Orosius* translation is particularly useful for our purposes because it includes some material that is not found in the Latin source, most famously the "Voyages of Ohthere and Wulfstan," two travelogues told by Old Norse speakers and recorded in the court of King Alfred.



lexomics.wheatoncollege.edu

(Slide 27)

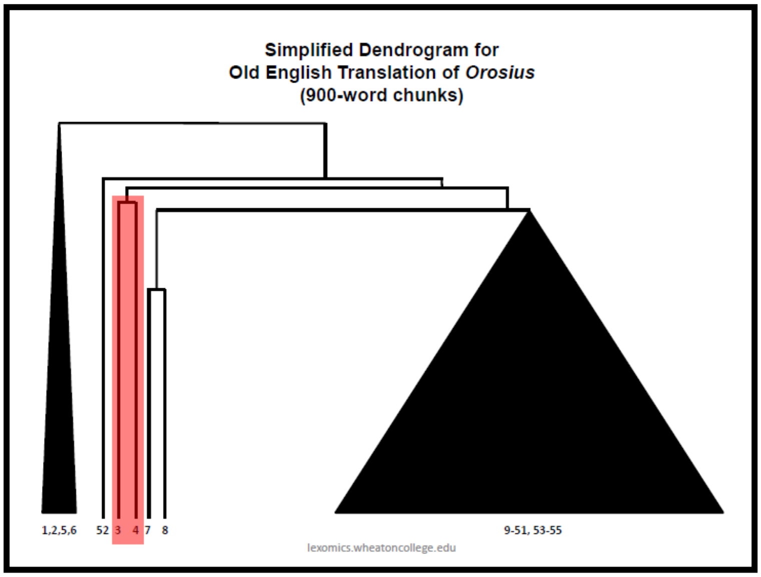
Because prose texts tend to be much longer than Anglo-Saxon poems, one challenge we face is that of interpreting very complex dendrograms, like this one.  To avoid becoming lost in a tangle of clades, we simplify the dendrogram by representing only the large-scale divisions.


Simplified Dendrogram for
Old English Translation of *Orosius*
(900-word chunks)

lexomics.wheatoncollege.edu

(Slide 28)

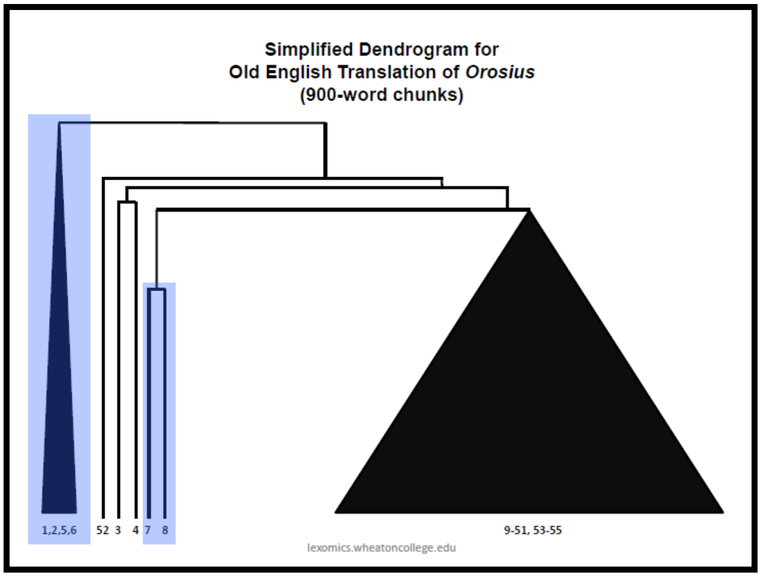
This simplified dendrogram shows that there are five main divisions in the *Orosius* text: a central clade, that is here represented by a large triangle, with a few smaller groupings that are different in vocabulary distribution. The central clade corresponds with material translated directly from the Latin text of Orosius' History.


Simplified Dendrogram for
Old English Translation of *Orosius*
(900-word chunks)

lexomics.wheatoncollege.edu

(Slide 29)

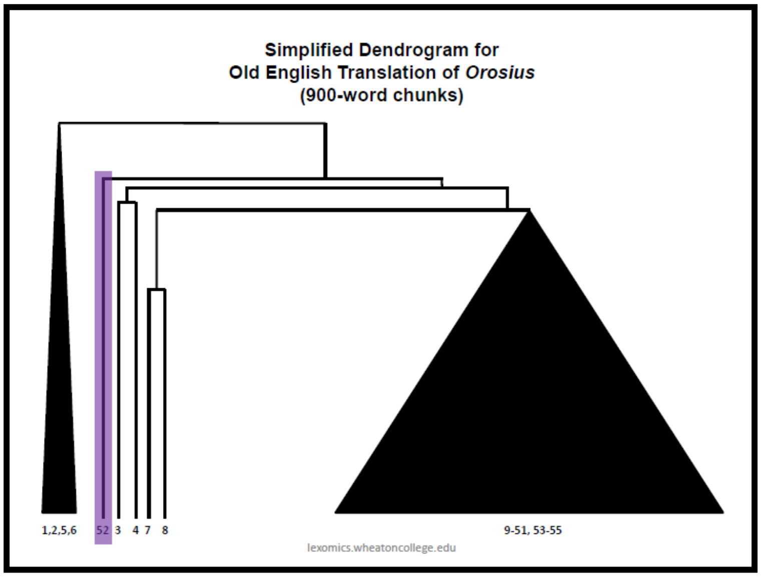
The "Voyages" passage, whose source is Old Norse rather than Latin, stands out rather clearly in chunks three and four.


Simplified Dendrogram for
Old English Translation of *Orosius*
(900-word chunks)
1,2,5,6
52
3
4
7
8
9-51, 53-55
lexomics.wheatoncollege.edu

(Slide 30)

Chunks one, two, five and six are in their own clade, separate from the rest of the text. This section of the Old English translation is a detailed discussion of the geography of the known world. *Orosius* has a geographical description in this same place, but the Anglo-Saxon text is a translation of some other geographical material, not Orosius' Latin. Orosius did not know about the northern and western parts of Europe, but these regions were of interest to King Alfred and his court.  The translator therefore used a different, more relevant source here. Chunks seven and eight are also geographical material, possibly from the same source and only separate from it in the dendrogram due to the interruption of the "Voyages."

We can therefore conclude that our dendrogram reflects the underlying source structure of the translation. The central clade is material from *Orosius*, chunks three and four from the Old Norse voyages, and chunks one, two and five to eight from an unknown geographical text or texts.


Simplified Dendrogram for
Old English Translation of *Orosius*
(900-word chunks)
1,2,5,6
52
3
4
7
8
9-51, 53-55
lexomics.wheatoncollege.edu

(Slide 31)

The source of chunk fifty-two is not known. But based on our previous analysis, we suspect that it either has a different source from the rest of the text, or is somehow stylistically distinct. We are seeking to understand the placement of chunk fifty-two and until we know why it is where it is in the dendrogram, we can simply note that Lexomics can be used to identify areas of a text that may have hitherto unidentified outside sources and would therefore repay additional close analysis.

Sometimes Lexomic methods confirm what we already know. Other times they help us to answer new questions. In still other cases—which are perhaps the most exciting—Lexomics just tells us where we need to look more closely.

By: Michael Drout and Leah Smith

**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."