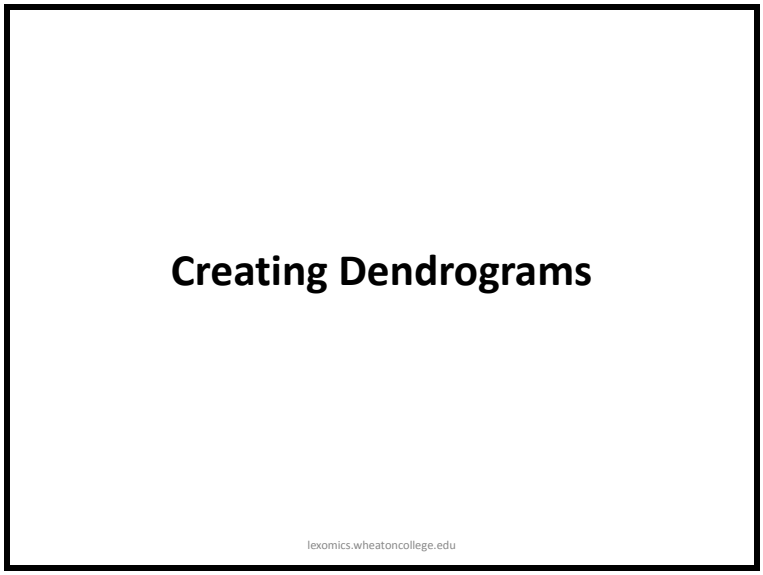
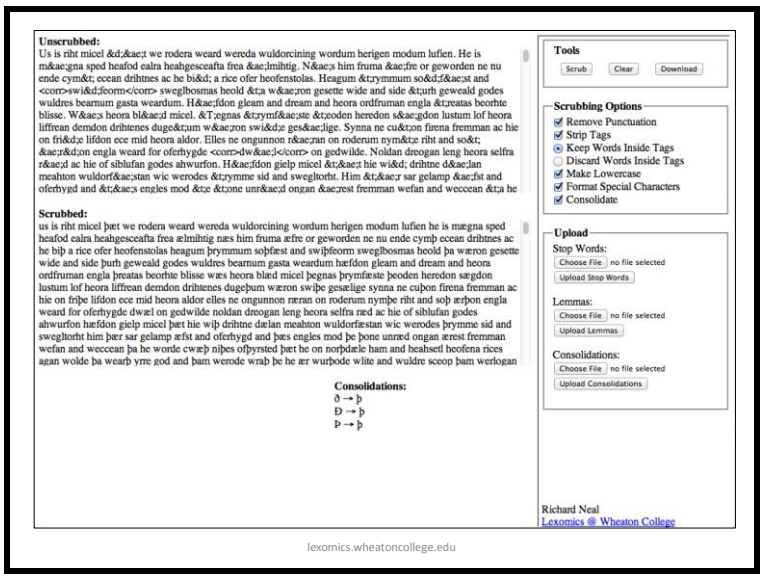
How to Create a Dendrogram



(Slide 1)

For our purposes, at the most basic level a dendrogram is a visual representation of word frequency in texts. This frequency can then be used to analyze the relationship between texts and their authors, sources, and other texts.

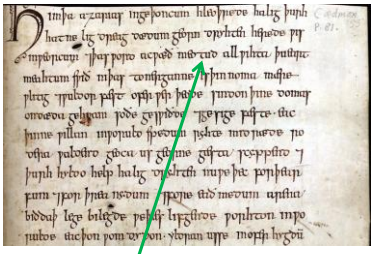


(Slide 2)

For Lexomic analysis, we begin with a procedure called “scrubbing,” in which we process the electronic texts to remove all formatting and punctuation and to replace all capital letters with lower-case, so that when we count the words we do not count “king” separately from “King.”

(Slide 3)

Word Count

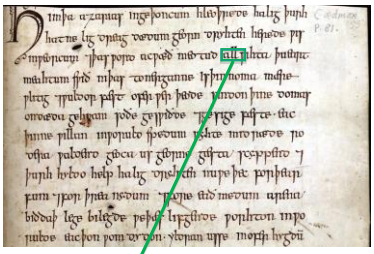


:	:
:	:
all	17
this	2
that	4
:	:

```
$wordCount{$word}++;
```


(Slide 4)

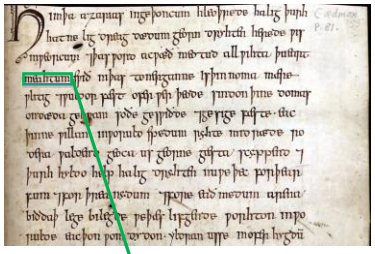
Word Count



:	:
:	:
all	18
this	2
that	4
:	:

```
$wordCount{$word}++;
```

Word Count



```
$wordCount{$word}++;
```

:	:
:	:
all	18
this	3
that	5
menigfeald	1
:	:

lexomics.wheatoncollege.edu 5

(Slide 5)

We then tabulate all the words in the entire text under consideration, cut the text into segments, and tabulate the words in each segment. We compute the relative frequencies of each word by dividing the number of times the word appears in a segment by the total number of words in that segment. With these frequencies calculated, we are ready to use statistical methods to compare the segments.

Let's look at a simplified example. Imagine we only have a two-word vocabulary and we want to compare the relative similarity of texts that use that vocabulary. Say we only have the words red and green.

Example: 4 Texts with 2-word Vocabularies

	Red	Green
Text 1	0	10
Text 2	5	1
Text 3	4	2
Text 4	6	4

lexomics.wheatoncollege.edu

(Slide 6)

We tabulate these words for each text. If the texts are different sizes, we need to make sure that we calculate the relative frequency of the word, dividing the counts by the size of the text or segment. The relative frequencies in our example would look like this.

(Slide 7)

Relative Frequency

	Red	Green
Text 1	0/10	10/10
Text 2	5/6	1/6
Text 3	4/6	2/6
Text 4	6/10	4/10

lexomics.wheatoncollege.edu

We can see from the chart that texts two and three have fairly similar vocabulary distributions. But we want a way to represent visually the similarity of all the texts. So we take these relative frequencies and use them to plot the texts according to each word.

(Slide 8)

Relative Frequency

	Red	Green
Text 1	0/10	10/10
Text 2	5/6	1/6
Text 3	4/6	2/6
Text 4	6/10	4/10

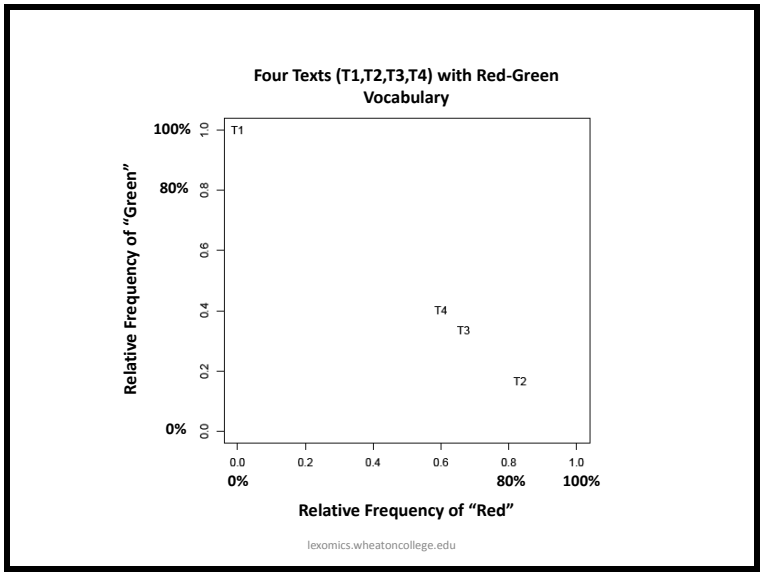
lexomics.wheatoncollege.edu

So in this example, we see that all of Text 1 is "green," which is why Text 1 is at the top left corner of the graph: 100% green and 0% red. When all the texts are plotted on the graph we can calculate the

distances between each point. These distances represent the similarities of the vocabularies of each text.

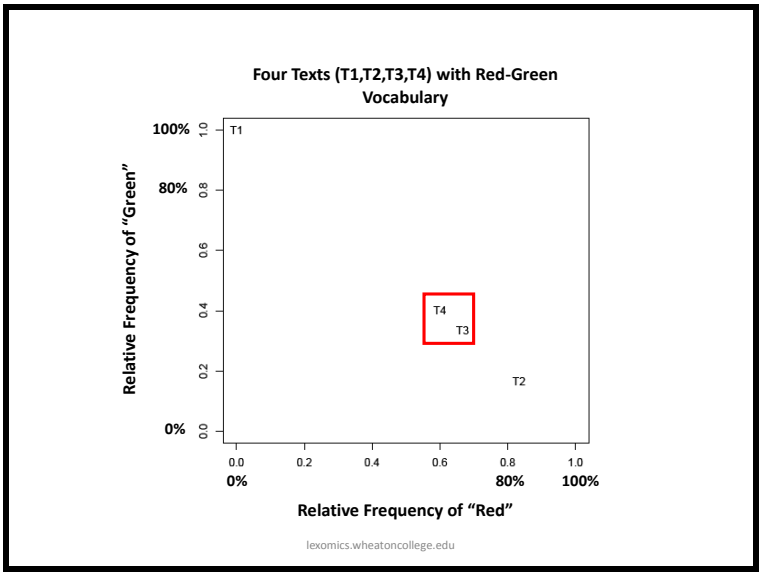
To convert this graph into a readable dendrogram, we use a statistical technique called hierarchical agglomerative clustering. To get an idea of what that means, let's create a dendrogram for our red-green example.

(Slide 9)



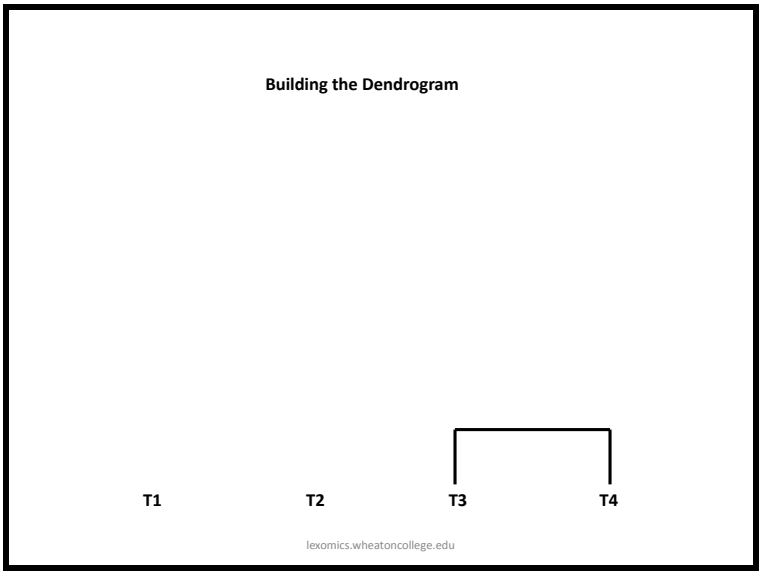
We start by determining which two texts are closest to each other on the graph by calculating all the distances and finding the smallest one. In our example, T3 and T4 are closest to each other, so we begin our tree diagram with these texts,

(Slide 10)



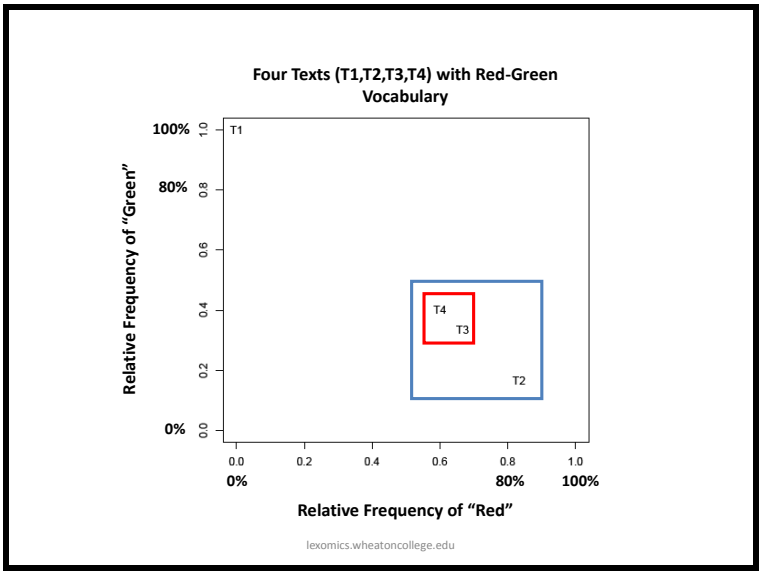
…placing them next to each other, and then using their distance from each other to determine the length of the lines between them and their connection point.

(Slide 11)

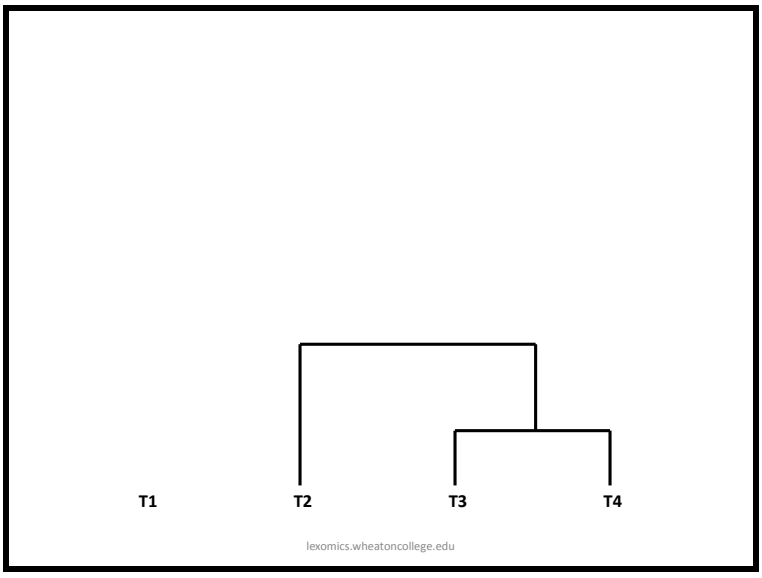


We now calculate the midpoint of the line between T3 and T4 and then find the text closest to this point. In this example, T2 is closest. We then calculate the distance between T2 and the average of T3 and T4 in order to determine the height of the next branch point on the dendrogram.

(Slide 12)

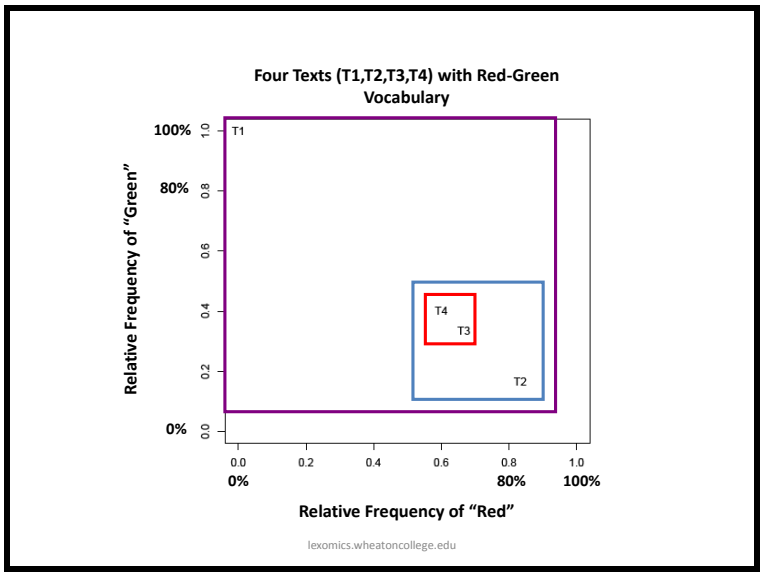


(Slide 13)



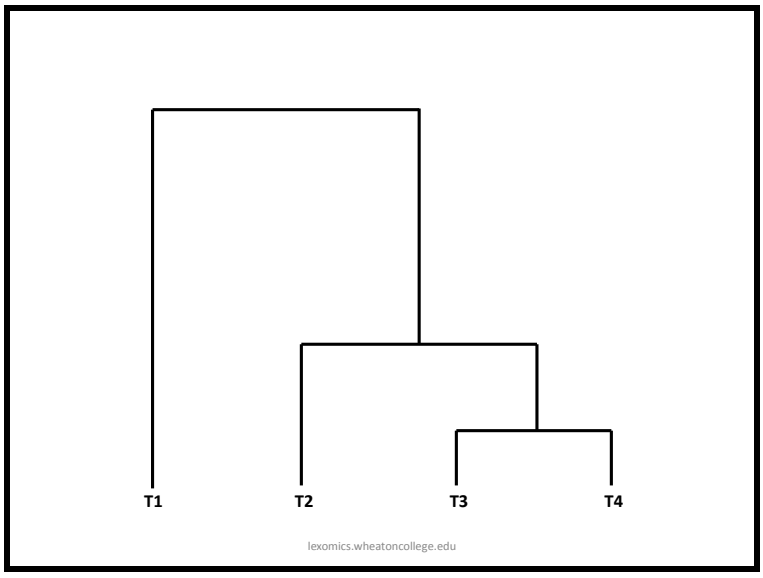
Repeat the process with the final text, treating T3, T4, and T2 as an averaged unit.

(Slide 14)



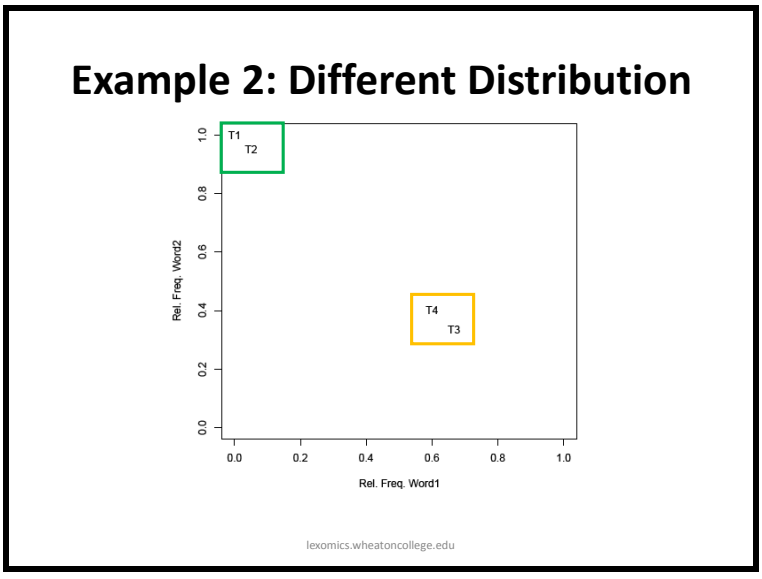
The resulting dendrogram would look like this.

(Slide 15)



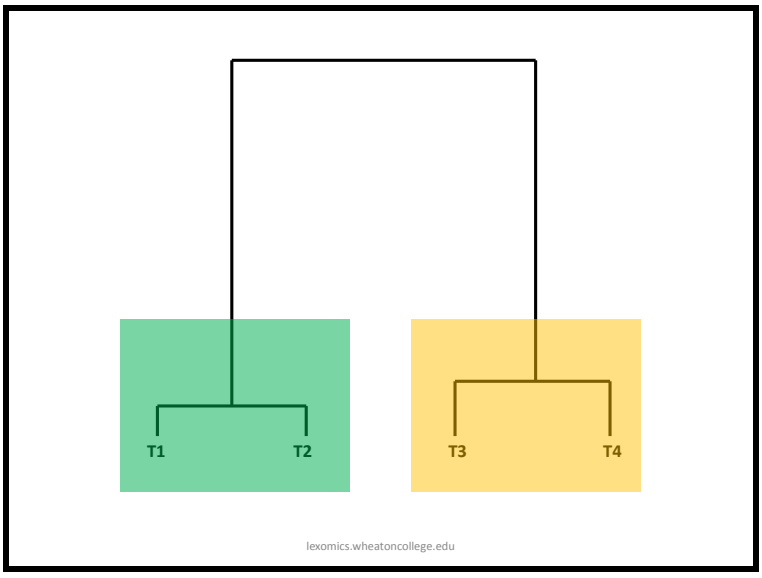
Dendrograms don't always have to have a stepwise geometry, with each leaf joining first with one, then with another, and so forth. Different patterns of word frequencies create different dendrograms, including sets of pairs, as in this dendrogram. Here, texts one and two are very similar to each other, as are texts three and four, but neither pair is particularly similar to the other pair.

(Slide 16)



Note that this dendrogram separates into two clades.

(Slide 17)



We have yet to come across many texts with only two words in them, but fortunately the same processes can be used for texts with thousands of words. When we compared red and green, we were using two dimensions. If we added blue, we would use a third. If we use 1000 words, we would need a thousand dimensions. And while you and I can't visualize that, a computer can be used to calculate the distances and use these to construct a dendrogram that represents the relationships of similarity and difference.

(Slide 18)

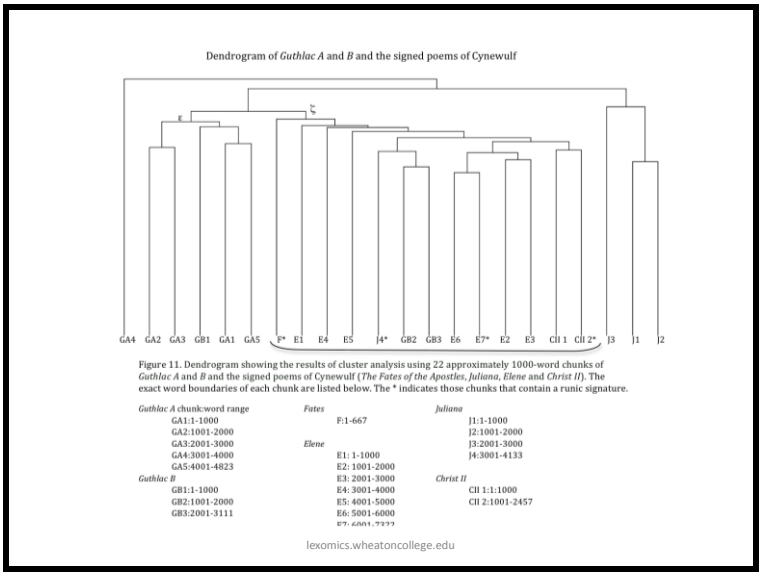


Figure 11. Dendrogram showing the results of cluster analysis using 22 approximately 1000-word chunks of *Guthlac A* and *B* and the signed poems of Cynewulf (*The Fates of the Apostles*, *Juliana*, *Elene* and *Christ II*). The exact word boundaries of each chunk are listed below. The * indicates those chunks that contain a runic signature.

Guthlac A chunk:word range	*Fates*	*Juliana*
GA1:1-1000	F:1-667	J1:1-1000
GA2:1001-2000		J2:1001-2000
GA3:2001-3000	*Elene*	J3:2001-3000
GA4:3001-4000	E1: 1-1000	J4:3001-4133
GA5:4001-4823	E2: 1001-2000	
Guthlac B	E3: 2001-3000	*Christ II*
GB1:1-1000	E4: 3001-4000	CII 1:1:1000
GB2:1001-2000	E5: 4001-5000	CII 2:1001-2457
GB3:2001-3111	E6: 5001-6000	
	E7: 6001-7322	

By: Michael Drout and Leah Smith

lexomics.wheatoncollege.edu

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