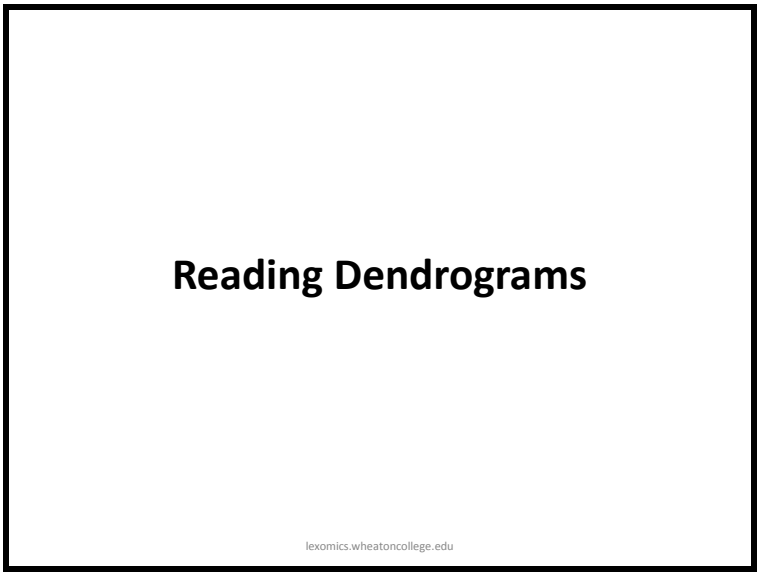
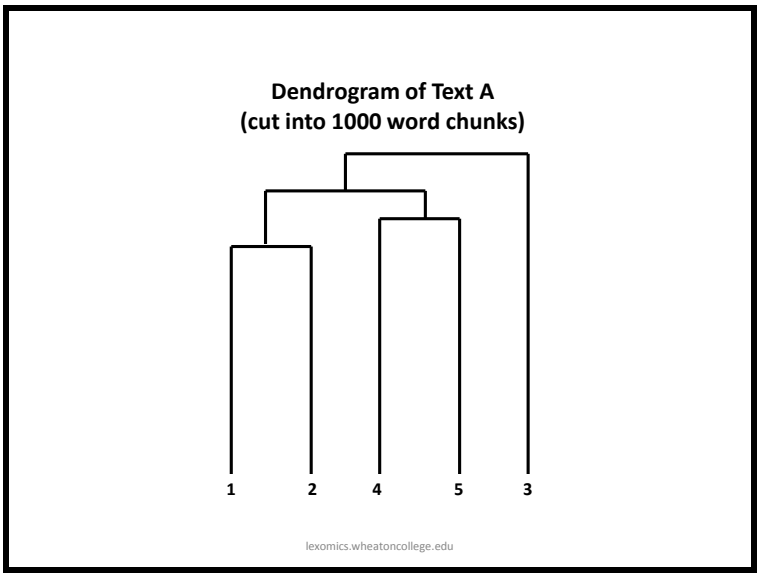
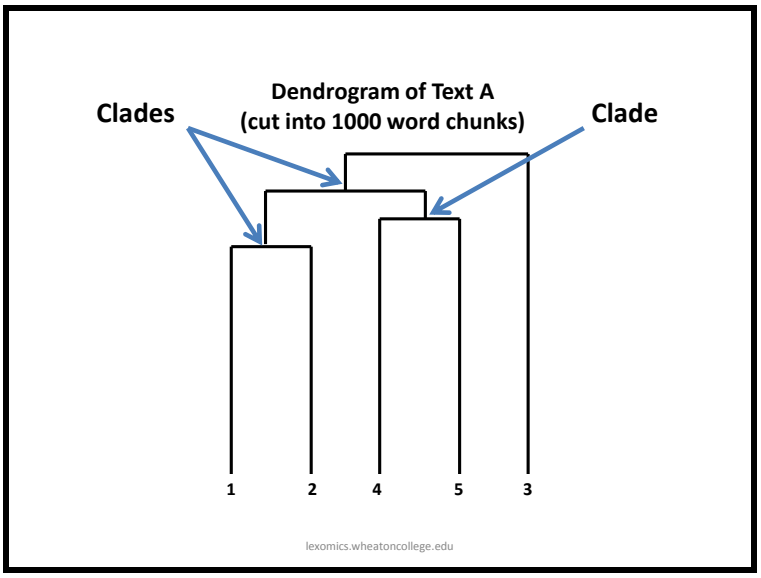
How to Read a Dendrogram

(Slide 1) 

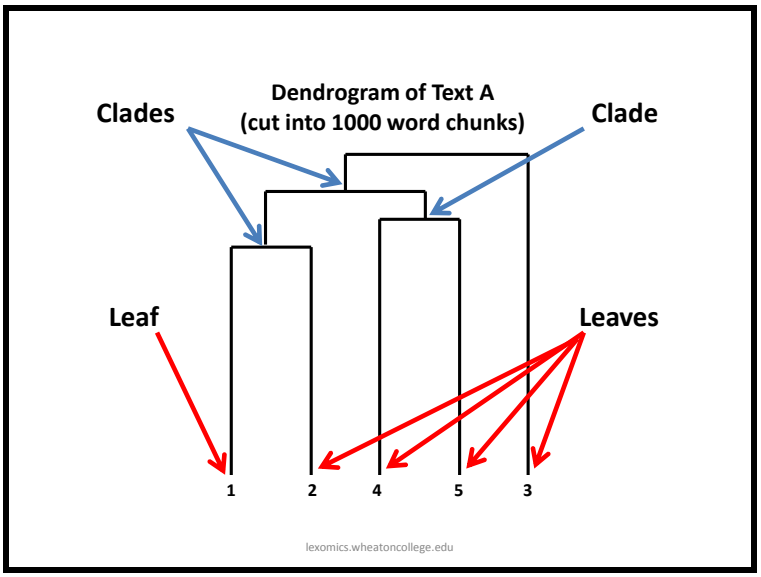
A dendrogram is a branching diagram that represents the relationships of similarity among a group of entities.

(Slide 2) 

Here we have a basic dendrogram.

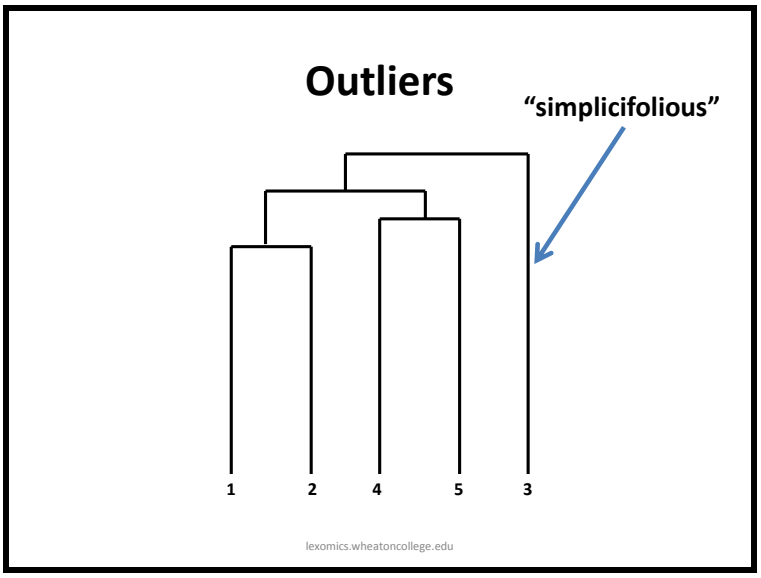
(Slide 3) 

Each branch is called a *clade*.

(Slide 4) 

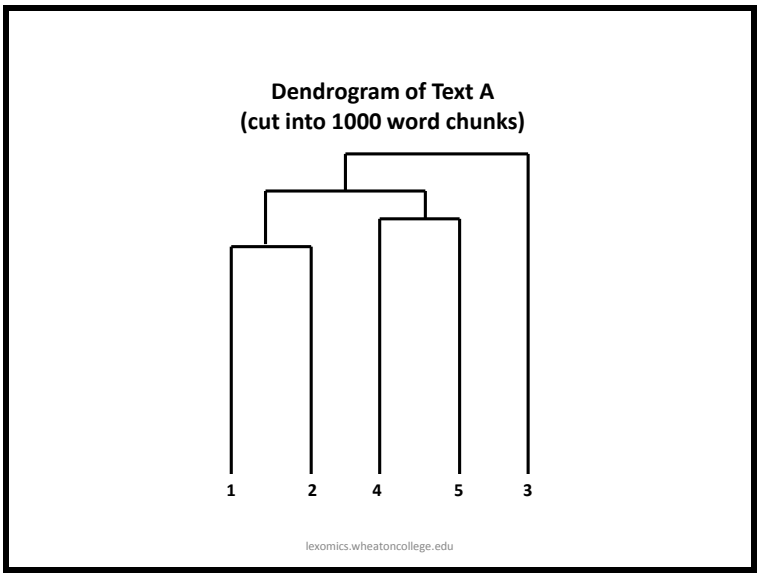
The terminal end of each clade is called a *leaf*.

(Slide 5)



Clades can have just one leaf (these are called *simplicifolious*, a term from botany that means “single-leafed”) or they can have more than one. Two-leaved clades are *bifolious*, three-leaved are *trifolious*, and so on. There is no limit to the number of leaves in a clade.

(Slide 6)



The *arrangement* of the clades tells us which leaves are most similar to each other. The *height* of the branch points indicates *how* similar or different they are from each other: the greater the height, the greater the difference.

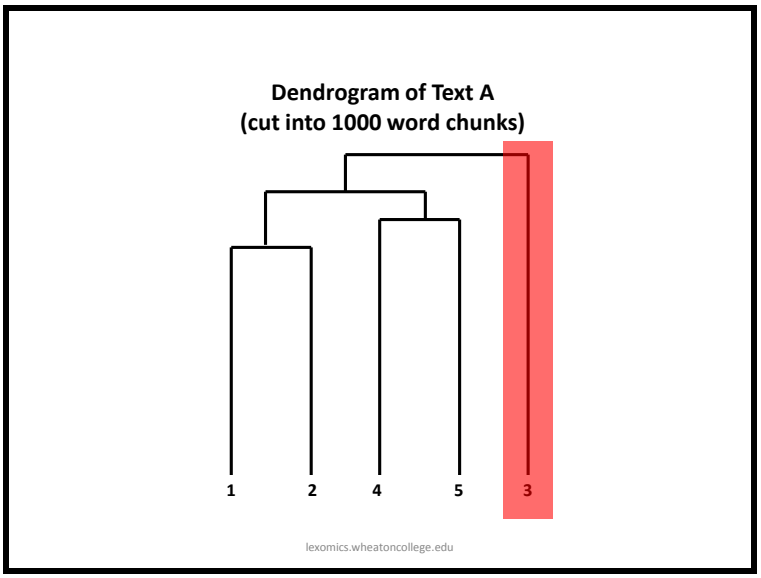
We can use a dendrogram to represent the relationships between any kinds of entities as long as we can measure their similarity to each other. In Lexomic analysis, we compare the distribution of different words among whole texts or segments of texts.

In this dendrogram, we have cut a text into 5 segments—also called chunks—that are each 1000 words long.

There are two ways to interpret a dendrogram: in terms of large-scale groups or in terms of similarities among individual chunks.

To identify large-scale groups, we start reading from the top down, finding the branch points that are at high levels in the structure.

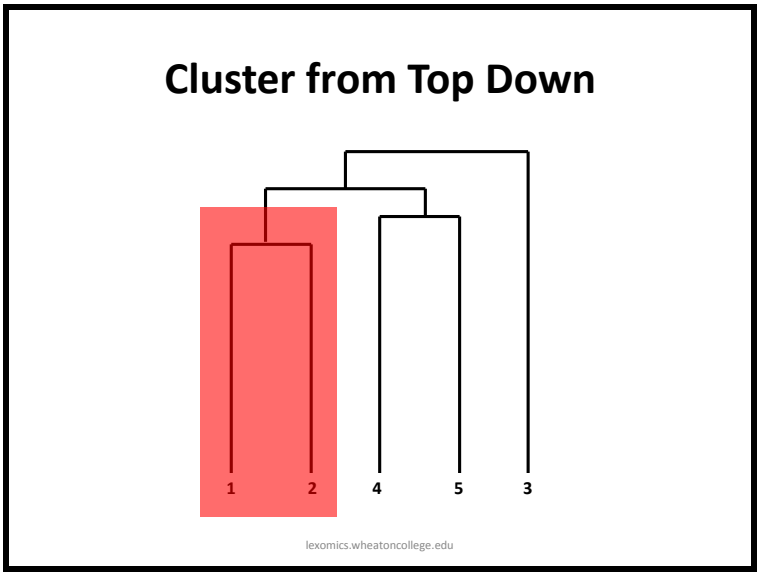
(Slide 7)

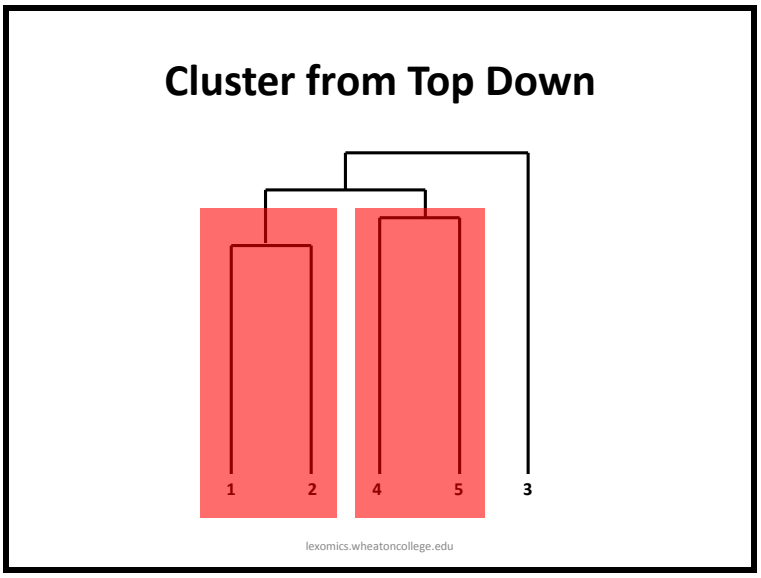


In this particular dendrogram, we see that chunk three is completely separate from all the others. Chunk three is therefore *simplicifolious*. We interpret its placement as indicating that the distribution of words in that chunk is substantially different from the distribution in the remaining chunks.

These other chunks can be clustered into groups.

(Slide 8)

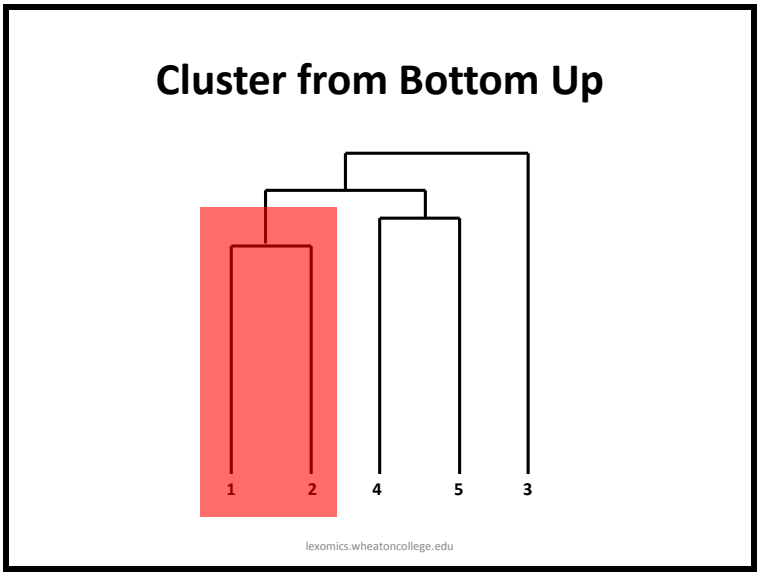




(Slide 9)

In this dendrogram chunks one and two are more similar to each other than they are to four or five. The branch containing chunks one and two is a clade. We usually label clades—at any level of the diagram—with Greek letters, moving from left to right and top to bottom.

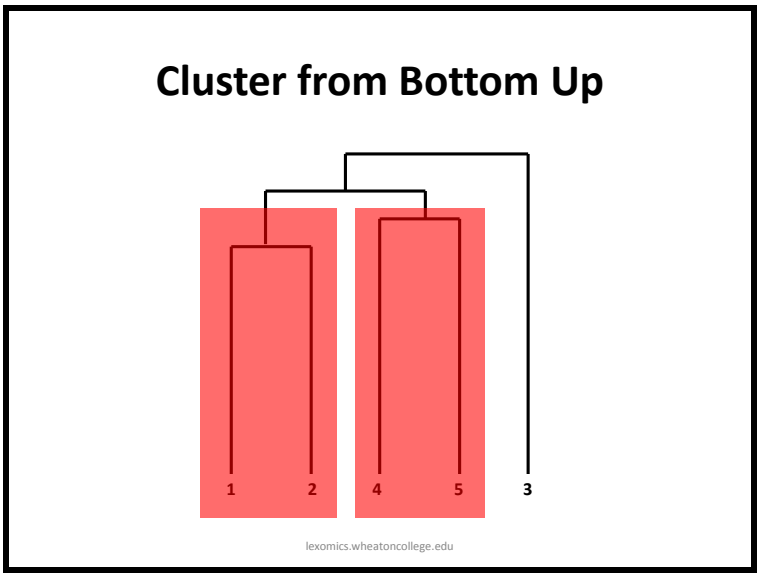
If we are trying to identify which individual segments are most similar to each other, we read the dendrogram from the bottom up, identifying the first clades to join together as we move from bottom to top.



(Slide 10)

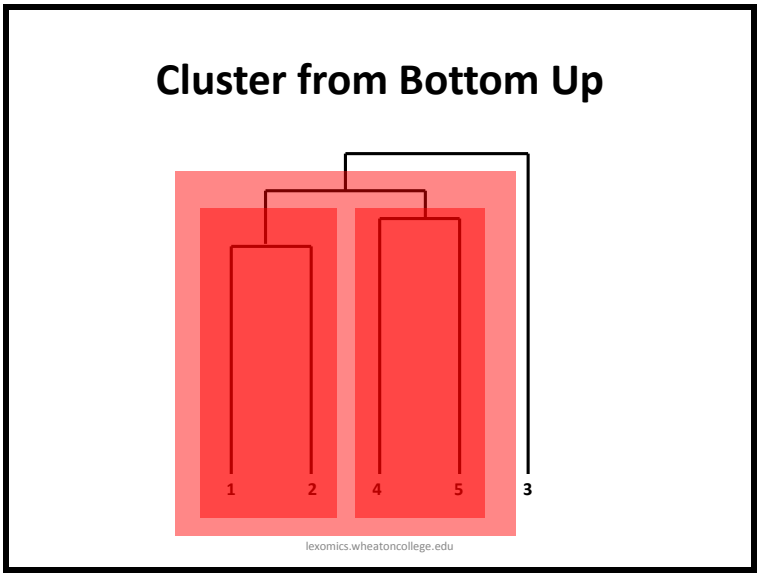
The connection between chunks one and two is the closest link to the bottom of the diagram. Therefore chunks one and two are most similar and join together first in the branching diagram.

(Slide 11)



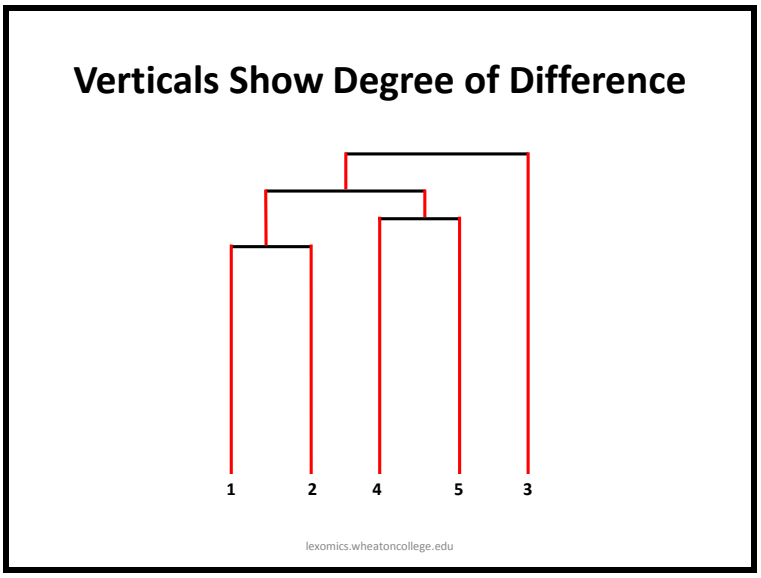
Chunks four and five are similarly clustered, indicating that chunks four and five are more similar to each other than they are to any other chunks.

(Slide 12)



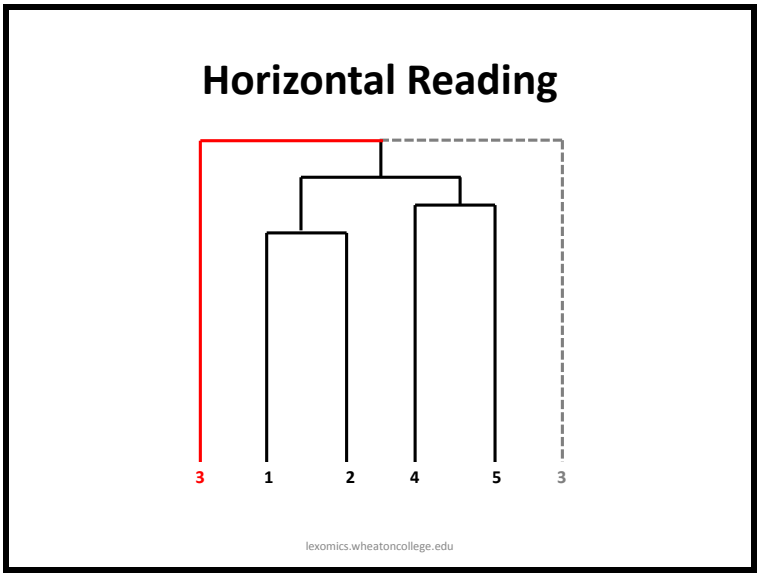
Moving up, we see that the next joining connects the clade with chunks one and two and the clade with chunks four and five. This geometry indicates that every chunk within that cluster is more similar to each other than to any chunks that join at a higher level (in this case the only other chunk is three).

(Slide 13)



The height of the vertical lines, highlighted here in red, indicates the degree of difference between branches. The longer the line, the greater the difference.

(Slide 14)



The horizontal orientation of dendrograms is irrelevant. Imagine the dendrogram as a mobile, in which the arms can shift position, but the vertical height and subgroup organization remain constant. For example, it makes no difference whether segment three lies on the left or the right of the other clusters.

By themselves, dendrograms only tell us a bit about the similarities of word patterns in texts. But when we link them to other types of information—like ribbon diagrams or traditional kinds of textual analysis—we can often draw significant conclusions. A dendrogram can tell you where to look more closely, and it can provide independent support or contradiction for various hypotheses about similarity and difference.

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."