1. It’s a pleasure to be giving my second talk at the CSNTM! The first time was two years ago, when I led a Q&A with interns about the CBGM. This year’s talk is more of a lecture. I will get into the CBGM in the second half, but the first half will be an introduction to the related subject of phylogenetics, which is the methodology I’m using for my PhD research. [0:25]

2. So let’s jump right in! We’ll start with some basic concepts that are common to both methods. [0:05]

3. When a textual tradition consists of multiple surviving witnesses, like manuscripts, the first thing we need is a basis for comparing them. The traditional approach is to line up their texts in a *collation*, highlighting their *variant readings* in segments of text called *variation units*. Here you can see Reuben Swanson’s collation of Luke, with the variation units highlighted. The witnesses are listed in the right margin. These units with numbers in parentheses are *constant*, meaning that all the witnesses agree at them. They’re usually not included in analyses, but they can help us estimate how often scribes and readers *didn’t* change the text. [0:45]

4. Powerful techniques from biology can be applied to textual criticism thanks to the insight that textual collations are analogous to DNA sequence alignments. The rows for species or taxa correspond to witnesses, the columns for sites or characteristics correspond to variation units, and the cells containing states like DNA nucleobases and gaps correspond to variant readings, including omissions. As with missing DNA data, lacunae and things like retroversions from other languages can be encoded as ambiguous states. [0:45]

5. This means that witnesses like manuscripts, at the most basic level, are just rows of the collation, or sequences of readings. But paratextual features of the physical artifacts themselves could be encoded as states in a similar way, and as I will explain later, the age ranges for witness can be incorporated in more complex phylogenetic methods. [0:30]

6. But we’ll get to that later. For now, I’ll start by explaining how phylogenetics can take witnesses encoded in this way and evaluate hypotheses about where they belong in textual history. [0:15]

7. The underlying model for textual history in phylogenetics is a family tree of the textual witnesses, commonly called a *stemma*. The family tree, as you’ll notice, is upside-down. This is because unlike people, who typically have two parents, a manuscript is assumed to be copied primarily from one source. The surviving witnesses to the text are located at the bottom of the tree, at the “leaves,” which are depicted here as solid circles with Latin letters. These witnesses are traced back along their branches to common ancestors that are no longer extant; these are referred to as “hypothetical ancestors” or *hyparchetypes*, and they are portrayed here as dashed circles with Greek letters. The appeal of phylogenetics is that it can reconstruct the readings of these lost ancestors. The earliest common ancestor of the surviving witnesses is called the *archetype*, and it is often denoted by a lowercase omega. It represents the earliest text that can be reconstructed from the surviving part of the tradition. If the tradition is well-sampled by its surviving witness, then this could correspond to the text as it entered circulation. But if entire early branches of the tradition have gone extinct, then this authorial text will correspond to a separate ancestor called the *root* of the stemma, denoted by a capital omega, and we may have to reconstruct it by conjecturally emending the text of the archetype. [01:30]

8. In short, a stemma represents a hypothesis about how witnesses are related and how close they are to the earliest text. The problem is that many different hypotheses are possible. In the examples here, the witnesses b1 and b2 could be siblings with equal textual value, or b2 alone could be a significant witness to a very early text. Our goal is to determine which hypotheses best explain the collation data that we have. And to do this, we need a way to measure how good a candidate stemma is. [0:30]

9. The traditional metric used in phylogenetic textual criticism is *parsimony*. Under this metric, the cost of a stemma is the smallest number of changed readings along its branches. This is basically the same idea as what the CBGM calls *coherence*. Both ideas are based on Ockham’s Razor: under the assumption that scribes copied faithfully more often than they erred or innovated, the hypothesis that requires the fewest violations of this assumption is best. We can conveniently calculate this cost for a given stemma at each variation unit independently and get the total cost by adding up these individual costs. Even more conveniently, there’s an efficient way to do this by starting from the leaves of the stemma and working our way up to the top.

For a demonstration of how this works, I’ll consider a single variation unit where five extant witnesses attest to common variations on a common name. [1:00]

10. The process of calculating a given stemma’s cost at a given variation unit consists of a forward pass and a backward pass. We only work up to the archetype, as this is the earliest ancestor whose reading can be reconstructed. We start with the lowest witnesses, c1 and c2. They have different readings, so we say that their common ancestor γ could have either of these readings. Next, we have the extant witness a and the hyparchetype γ: a reads χριστου ιησου, and γ also has this as one of its potential readings, so their common ancestor α reads χριστου ιησου. On the same level, we also have b1 and b2: they agree on the reading ιησου χριστου, so their common ancestor β also has this reading. That means that we have a tie at the archetype: it could have either of these readings. [Click to next slide]

The backward pass starts from the archetype and goes down to the leaves. We can pick any reading that the archetype could have based on the forward pass. If we choose χριστου ιησου and then resolve the ambiguities from the forward pass as we go down the branches, we get these ancestral readings. The stemma therefore features two changes in reading, resulting in a cost of 2. [Click to next slide]

If instead we choose ιησου χριστου as the reading of the archetype, then we get these ancestral readings, but the cost of the stemma remains the same. [1:30]

11. Now let’s see what cost we get for a different stemma. We’ll work through the forward pass again. Like before, c1 and c2 have different readings, so their ancestor γ could have either of these readings. At the next level, we see that γ and b1 also have no readings in common, so their ancestor β could have one of three readings. We finally can break the tie once we get to the next level: a and β have the reading χριστου ιησου in common, so we set this as the reading of their ancestor α. Finally, we have another tie at the hyparchetype between the reading of α and the reading of b2. [Click to next slide]

We can start with either reading in the backward pass. If we start with χριστου ιησου, then we get these ancestral readings down the branches, with three changes of reading in total. [Click to next slide]

And if we start with ιησου χριστου, then we get these ancestral readings, also with a cost of 3. [1:00]

12. In fact, it’s been proven that if the cost of a stemma is calculated this way, it will be the same no matter how we break ties at the archetype. Even more importantly, the cost of the stemma will be the same *no matter where the archetype and root are located*! This handy fact has enabled textual critics to separate their work into an automated part and a human part: the computer can calculate the costs of millions of stemmata without needing to know where their roots are, and afterwards, the textual critic can identify where the root belongs based on human judgments about the internal evidence of readings. This, in fact, is how computer-assisted textual criticism has traditionally been done. [0:45]

13. But what if we want to incorporate internal evidence of readings into the calculation of a stemma’s cost? One benefit of doing this is that we get a clean separation of concerns between different types of internal evidence. There are two main types of internal evidence: intrinsic probabilities, which concern what the author most likely wrote, and transcriptional probabilities, which concern what later scribes were most likely to do. If we just lump these two types of evidence together to determine where the root of the stemma is, it’s not clear how we’re supposed to decide between them when they point in opposite directions. But if we introduce intrinsic and transcriptional evidence earlier in the process, then we can use them in distinct ways. With intrinsic evidence, we can include the root in the stemma (like we have in this picture) and specify which reading or readings it could have. This will affect how we calculate the cost of a stemma in the backward pass. With transcriptional evidence, some types of scribal changes—like skips of the eye, confusions involving similar sounds, and harmonizations—may be more likely than others, and a change from one reading to another may not be as easy to make in the opposite direction. We’d like for this to figure into our calculation of a stemma’s cost.

The cool thing is that we can extend the traditional approach to do this. The main trade-off, of course, is that the cost of the stemma will now depend on where its root is. [1:30]

14. Let’s start by incorporating just intrinsic evidence. We’ll pretend that we’re being particularly bold and we’ve conjectured that the author originally wrote ιησου, a reading not attested in the surviving tradition. Then this would be the reading at the root, and the forward pass up to the archetype would go as it normally would. [Click to next slide]

But in the backward pass, we’ll have an extra change of reading, and there are multiple ways to explain it. One possibility is that an early scribe expanded the author’s ιησου to χριστου ιησου, and the scribe behind the β branch then transposed this to ιησου χριστου. [Click to next slide]

Another possibility is that an early scribe expanded ιησου to ιησου χριστου, which the scribe behind the α branch transposed to χριστου ιησου. [Click to the next slide]

The third possibility is probably the most transcriptionally plausible: two branches of the tradition independently expanded ιησου in different ways, with the α branch adding χριστου before it and the β branch adding χριστου after it. [1:15]

15. We can compare scenarios like these in a quantifiable way if we assign different costs to different types of changes between readings. It’s helpful to think of this in terms of a graph like the one on the right: the nodes are the readings, and the edges between them correspond to changes with different costs. It’s a way of modeling the transcriptional behavior of the average scribe. The costs on the edges can be conveniently encoded in a table called a *cost matrix*, which is pictured to the left here. The starting readings are along the rows, and the resulting readings are along the columns, so the cost of starting at one reading and changing to another is located at the cell in that row and column. Naturally, the cost of copying the same reading faithfully is just 0. A skip of the eye, which would explain the loss of χριστου after ιησου or the loss of ιησου after χριστου, would be very common, so we’ll assign it a low cost, say, 2. An expansion of ιησου or χριστου to ιησου χριστου or χριστου ιησου would be almost as common, so we’ll assign it a slightly higher cost at 3. Substitutions between ιησου and χριστου and transpositions between ιησου χριστου and χριστου ιησου are even less common, so we’ll assign all of them a cost of 5. Finally, for the changes we can’t explain, like the shortening of χριστου ιησου to ιησου and the shortening of ιησου χριστου to χριστου, we’ll assign a high cost, like 10. [1:30]

16. We can then use the cost matrix to calculate the cost of a stemma while taking all transcriptional possibilities into consideration. This is a bit more involved than the forward and backward passes we did earlier, so I’ll walk you through this one from the start. The first thing you’ll notice is that we don’t just keep track of which reading or readings the witnesses and their ancestors have; instead, for every witness or ancestor, we keep track of its lowest total cost for every reading it could have. Thus, every witness or ancestor has four entries for costs—one for each of the four variant readings in this unit. Every extant witness has a cost of 0 for its known reading and infinity for every reading it is known not to have.

So let’s figure out what the cost entries for γ should be based on c1 and c2. If γ reads ιησου, the first reading, then its minimum cost would be the cost of ιησου changing to χριστου in c1 plus the cost of ιησου being expanded to χριστου ιησου in c2. Checking the cost matrix, we see that these costs are 5 and 3, respectively. We would then add these values to the costs of c1 reading χριστου and c2 reading χριστου ιησου, but right now, those values are 0, so we can ignore them. Adding up the transition costs of 5 and 3 from the matrix gives us a total cost of 8 for γ reading ιησου. You can work out the other entries for γ as an exercise. [1:45]

17. Now let’s go up a level and calculate the costs for α. If α reads ιησου, then we have a cost of 3 for it expanding to χριστου ιησου in witness a, plus 0, since a has a cost of 0 for that reading. But what about the cost that γ contributes? To calculate this, we have to consider the cost of each reading that γ could have, plus the cost of ιησου changing to that reading, and we pick the smallest of these costs. So, if γ reads ιησου, then we add its cost of 8 to the cost of 0 for α’s reading not changing. If γ reads χριστου ιησου, then we add its cost of 2 to the cost of 3 for ιησου being expanded to it. If γ reads ιησου χριστου, then we add its cost of 15 to the cost of 3 for ιησου being expanded to it. And if γ reads χριστου, then we add its cost of 3 to the cost of 5 for ιησου being changed to χριστου. The lowest of these costs is 5, which occurs if γ reads χριστου ιησου. So we add this cost of 5 to the cost of 3 for ιησου changing to χριστου ιησου in a, and we get a total cost of 8.

Now suppose that α reads χριστου ιησου. Then it would have a total cost of 0 along the branch to a, because their readings would agree. That just leaves the cost along the branch to γ. If γ reads ιησου, then we add its cost of 8 to the cost of 10 for χριστου ιησου changing to it. If γ reads χριστου ιησου, then we add its cost of 2 to the cost of 0 for χριστου ιησου remaining unchanged. If γ reads ιησου χριστου, then we add its cost of 15 to the cost of 5 for χριστου ιησου being transposed to it. And if γ reads χριστου, then we add its cost of 3 to the cost of 2 for χριστου ιησου being shortened to χριστου. The smallest of these costs is 2, so the cost of α reading χριστου ιησου is 0 + 2, or 2. [2:30]

18. If, on the basis of intrinsic evidence, we narrow the reading of the root down to one or more options, then can we use the same process to determine the cost of the stemma from the root. So if the root reading is ιησου, then the smallest cost is achieved if the archetype also reads ιησου: the cost of the archetype reading ιησου is 8, and the cost of the reading staying the same from the root to the archetype is 0, resulting in a total cost of 8.

But, if the root reading is χριστου ιησου, then the smallest cost is achieved if the archetype also reads χριστου ιησου: the cost of the archetype reading χριστου ιησου is 7, and the cost of the reading staying the same is 0, resulting in a total cost of 7.

This also means that if we were undecided between these two readings for the authorial text, χριστου ιησου would be preferable, because it results in a slightly lower cost for the stemma. [1:00]

19. The use of variable costs is in fact a stepping stone to an even more substantial improvement: we could incorporate intrinsic and transcriptional evidence as probabilities between 0 and 1, and in this way, we could evaluate stemmata based on their probabilities rather than their costs. Based on intrinsic evidence, the root could have different probabilities for different readings based on how well they fit the author’s argument and style. Likewise, transitions between different readings could be assigned probabilities based on how likely scribes were to make certain types of changes. Probabilistic models of states and transitions between them are well-understood in mathematics, where they’re called *Markov chains*. We don’t have time to get into the technical details of this, but it should suffice to say that phylogenetic methods that work with probabilities have been developed and used successfully in biology and in textual criticism, including that of the New Testament. [1:00]

20. As a final note on phylogenetics, it is worth mentioning that when we work in probabilities, we can estimate various other parameters of interest while we evaluate stemmata. If we don’t know how likely different types of scribal changes are up front, we can estimate their probabilities as parameters in our transmissional model. This means that we can estimate average scribal habits automatically. Here you can see a histogram approximating how the rates for common classes of scribal changes are distributed. We can also use branches of varying lengths to model more copying events or copyists who made more errors on average. We can also use *clock models* so that these branch lengths can reflect durations of time. These clock models allow us to include information about the date ranges of witnesses, too, which is something traditional phylogenetics and other methods can’t do. Perhaps most importantly, working with probabilities allows us to quantify how certain we can be about different hypotheses about the history of the text. While working with all of these additional parameters is much more complex than the basic forward and backward pass I described at the start, it makes for a much more robust model, and modern computers are powerful enough to handle the complexity. [1:30]

[21 minutes for the phylogenetics half]

21. That brings us to the second half of today’s lecture, which concerns the Coherence-Based Genealogical Method, or CBGM. Once again, we’ll start with the basic ideas behind the method, and then we’ll take a closer look under the hood. As we do this, we’ll note comparisons to phylogenetics where they come up. [0:15]

22. The CBGM was developed over the course of two decades by Gerd Mink. One of his goals was to find a way to manage the problem of *contamination*, or mixture of sources. The idea is illustrated here: the witness b1 inherits different readings from the two β ancestors in different branches. This is common in the tradition of the New Testament. In fact, New Testament textual critics have considered phylogenetic methods unusable for this very reason. [0:30]

23. Gerd Mink based the CBGM on four fundamental principles or operating assumptions about textual transmission. The first one here is just the assumption that faithful copying is more likely than error or innovation, as I mentioned a while back. The second basically states that variant readings that aren’t original don’t arise out of thin air; they have to be explained from other readings. The third is that scribes typically used fewer sources rather than many. Intuitively, we’d expect most scribes to do what is easiest and copy from just one exemplar. There were surely readers and emendators who did the extra work of consulting other sources or making corrections against them, but they were the exception rather than the rule. The fourth is that scribes typically used closely related sources rather than distant ones. This would also require less effort, and most manuscripts in the same scriptorium or monastery were likely of a similar character. Certainly, manuscripts of diverse geographic origin and textual character traveled great distances and sometimes did end up together, but again, this was more exceptional than normal. All of these assumptions are also suitable for phylogenetics, although the last two don’t come into play as much, because traditional phylogenetics doesn’t model contamination.

There are a couple other functional assumptions that the CBGM makes. One is that witnesses represent abstract states of the text rather than physical artifacts bound by age and other features. The CBGM must stress this point because it sometimes has to treat younger witnesses as ancestors to older ones. The point is that the relationship of ancestry is not between two manuscripts—which would be impossible due to their ages—but between a “better” text and a “worse” text.

Such situations can arise because of a more crucial assumption the CBGM makes—namely, that there are no hypothetical ancestors allowed. The only exception is the initial text, also called the *Ausgangstext*. In phylogenetic terms, the CBGM doesn’t work with hyparchetypes; it just keeps the earliest text at the root of the stemma and uses the extant witnesses at the leaves as substitutes for the lost parts in the middle. The justification for this is that contamination makes the construction of a stemma, including hyparchetypes, impossible. But this comes with a functional advantage: the problem becomes simpler, because we now only have to relate witnesses whose readings are fully known. [3:00]

24.