Horses or farmers? The tower of Babel and confidence in trees

Geoff Nicholls, 2008.

Presenter: Sarah Schieferstein

Historical Linguistics

- Cognates
  - Same meaning
  - Same common ancestor
  - NOT a borrowing or accident
- Certain sound changes are more common (laziness)
- Can represent languages' evolution as a tree
  - Languages undergo evolution AKA sound change and word loss/birth
  - More cognates == closer in the tree

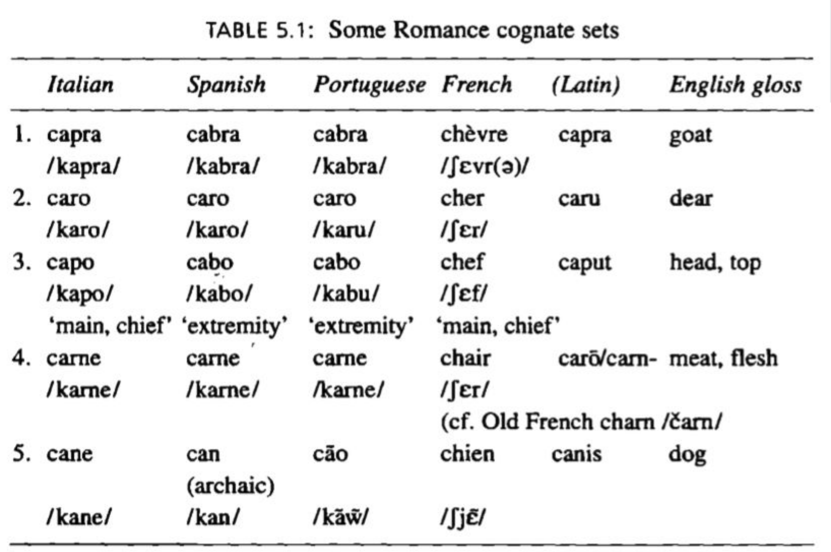
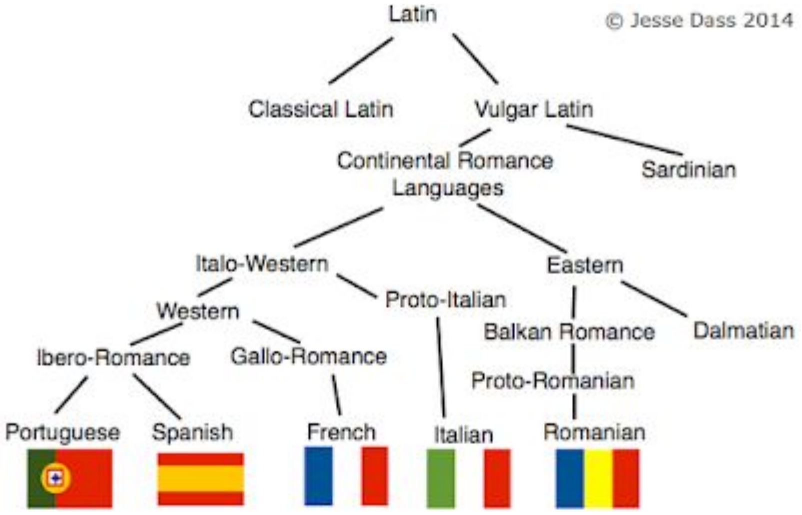


TABLE 5.1: Some Romance cognate sets

	Italian	*Spanish*	*Portuguese*	*French*	*(Latin)*	*English gloss*
1.	capra /kapra/	cabra /kabra/	cabra /kabra/	chèvre /ʃɛvr(ə)/	capra	goat
2.	caro /karo/	caro /karo/	caro /karu/	cher /ʃɛr/	caru	dear
3.	capo /kapo/ 'main, chief'	cabo /kabo/ 'extremity'	cabo /kabu/ 'extremity'	chef /ʃɛf/ 'main, chief'	caput	head, top
4.	carne /karne/	carne /karne/	carne /karne/	chair /ʃɛr/ (cf. Old French charn /čarn/	carō/carn-	meat, flesh
5.	cane /kane/	can (archaic) /kan/	cão /kãw̃/	chien /ʃjɛ̃/	canis	dog



© Jesse Dass 2014

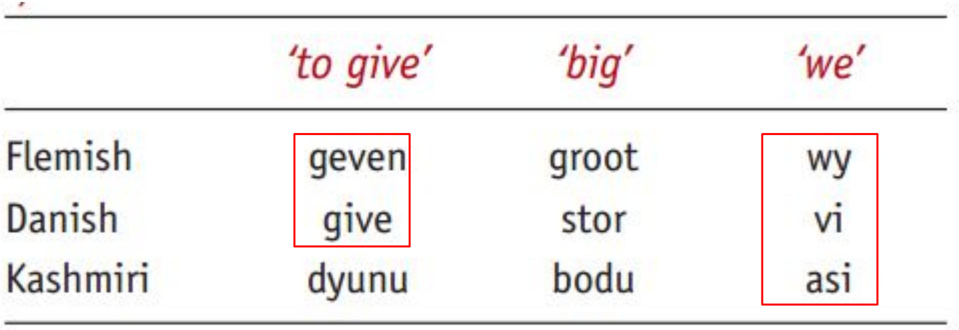
Indo-European’s root: horses or farmers?

- Did Neolithic farmers spread proto-Indo-European? Or did the Kurgan horsemen?
    - Farming begins: 8500 years ago
    - Horse ownership: 6500 years ago *(more likely via archaeology)*
- Using cognates, we can attempt to date the splits in the language tree
- Naive dating attempt: glottochronology. Find t = time separating 2 languages
    - μ = mean word lifetime (issues?), $n_1 + n_2$ = #cognates, n_{12} = #shared cognates

$$\hat{t} = \hat{\mu} \log \left(\frac{n_1 + n_2}{2n_{12}} \right)$$

Beyond glottochronology: use DNA models?

- Gray and Atkinson (2003)
  - Presence / absence of cognate
  - Use Bayesian DNA software MrBayes
  - Allow word lifetime rate variation
  - Use calibration points (well-known splits) to learn actual word lifetime rates, constrain scale of heterogeneity rate difference
  - **Confidence intervals**



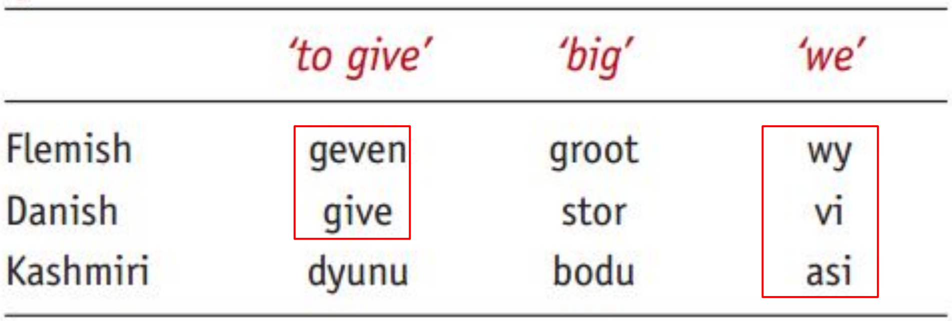
	'to give'	'big'	'we'
Flemish	geven	groot	wy
Danish	give	stor	vi
Kashmiri	dyunu	bodu	asi

To

$$X = \left(\begin{array}{cc|ccc|c} 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{array}\right)$$

Beyond glottochronology: use DNA models?

- Atkinson et al. (2005)
  - Different dataset
  - New model
  - Homoplasy free (like cognates)
  - Parameters: cognate birth, loss, split rate
- Both have Indo-European's root near 8500, not 6500! This disagrees with archaeological evidence!



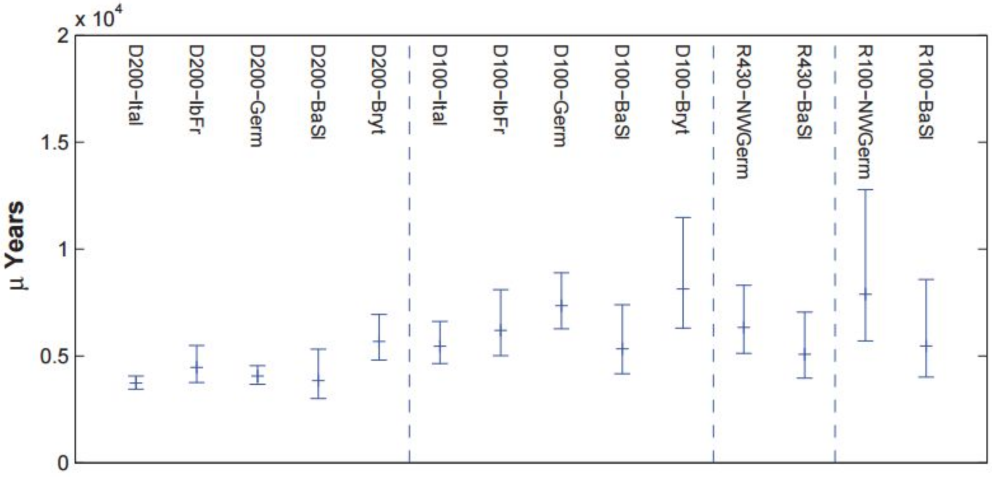
	'to give'	*'big'*	*'we'*
Flemish	geven	groot	wy
Danish	give	stor	vi
Kashmiri	dyunu	bodu	asi

To

$$X = \left(\begin{array}{cc|ccc|c} 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \end{array}\right)$$

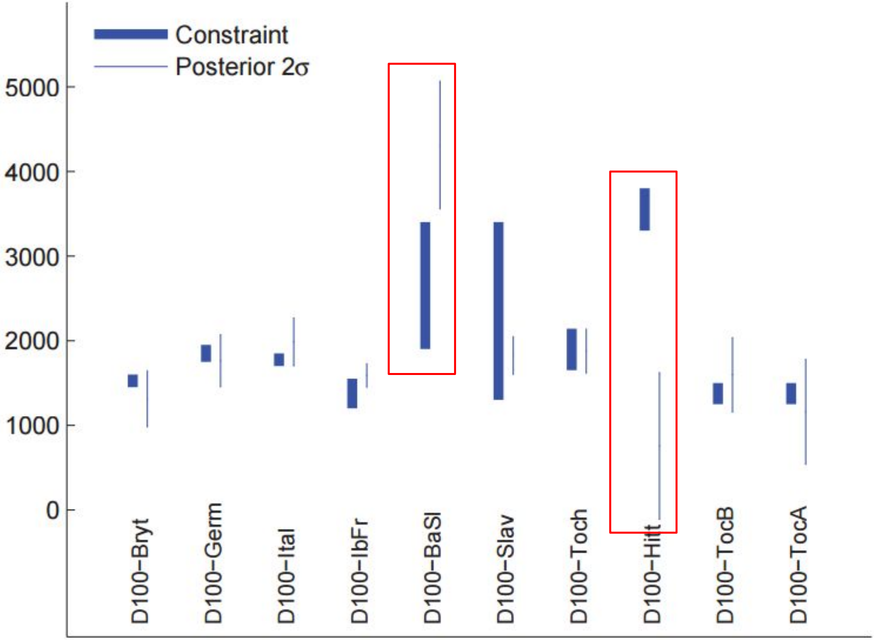
Checking the model: ‘known’ avg. word lifetime

- Model could be wrong if variation in μ, word lifetime per language group, is so high the model can’t reconstruct it
- To the right: estimated mean lifetime of cognates for known language splits
  - It’s low
  - Does not account for the root being 2000+ years off



Cross-validating the model

- Estimate the age of a language group without knowing its true split. Does it match the expert truth?
- Usually…
  - Thinks Hittite is more recent
  - Thinks Baltoslav is older
  - Baltoslav’s expert truth is based only on archaeological data
  - Hittite is missing many cognates but they are marked as absent



Thoughts

- The cognates should not be represented as binary features when they are naturally categorical
    - Assumes independence among characters, which is untrue
- But if the input is not binary, how does the input maintain polymorphism (e.g. a language has multiple cognates?)
- Are cognates the best way to represent the language tree? (What about languages with heavy borrowing or multiple parents across a tree, like creole languages?)
- Author mentioned: There is no distinction between an absent cognate and an unobserved cognate (missing data)
- Author mentioned: In general, the data seems very sparse and unreliable. There is missing data that isn't handled properly, the cognate classes are hand-made and doubtful, and some calibration points are questionable and wide-ranged

Takeaways

- Confidence intervals allow a thorough exploration of the data
- Calibration points allow methods to learn word lifetime rates and date properly by constraining the heterogeneity of possible trees
- The models and data are not flawless, but they show that there is repeatable statistical uncertainty in the date of Indo-European's root
- Most linguists side with the Kurgan/horse hypothesis still yet due to existing archaeological research; statistical explorations do not get the 'big picture'

References

- Figures in slide 2 from https://www.britannica.com/topic/Romance-languages/Linguistic-characteristics-of-the-Romance-languages and http://jessepaedia.blogspot.com/2014/04/what-living-language-is-closest-to-latin.html top to bottom
- All other figures from Nicholls 2008
- Nicholls, Geoff. "Horses or farmers? The tower of Babel and confidence in trees." *Significance* 5.3 (2008): 112-117.
- Atkinson, Quentin, et al. "From words to dates: water into wine, mathemagic or phylogenetic inference?." *Transactions of the Philological Society* 103.2 (2005): 193-219.
- Gray, Russell D., and Quentin D. Atkinson. "Language-tree divergence times support the Anatolian theory of Indo-European origin." *Nature* 426.6965 (2003): 435.