

On dating Proto-Indoeuropean via Bayesian phylogenetic inference

Igor Yanovich¹, Armin Buch, Johannes Dellert, Marisa Köllner, Fabrício Marcel Ferraz Gerardi, Roland Mühlenbernd, Johannes Wahle and Gerhard Jäger²

Institute of Linguistics, Tübingen University

¹Carnegie Mellon

²Netherlands Institute for Advanced Study

MPI for the Science of Human History, Jena

October 13, 2015



Bouckaert et al:

- based (mostly) on written record
- upper and lower limit: confined to last two millenia
- only lower limit: go up to last four millenia

this study

- derive upper and lower limits from archaeological findings

Proto-Indo-Iranian

Indo-Iranian

- Identification of Proto-II with Andronovo culture (Kuz'ima 2007)
- accepted by (some) proponents both of Anatolian and of Steppe theory
- time constraint: 1900–1600 BCE

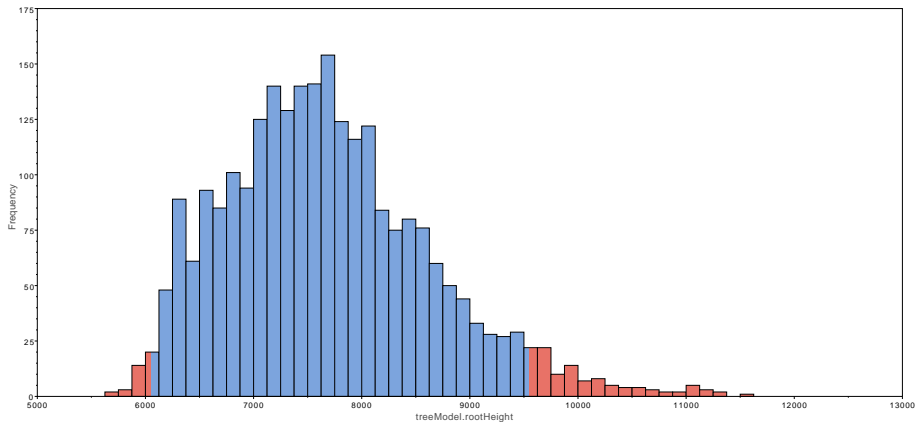
“Following the dissolution of the Balkan PIE linguistic area it seems likely that there was some further and significant punctuation episode in the Pontic steppe area which motivated the eastern dispersal of early Indo-Iranian into areas where the Early Steppe form of (Proto) IE had already been spoken for some time. Early Indo-Iranian is often correlated with the Andronovo culture of the Eurasian steppes, seen after 2000 BC (Kuzmina 1994)” Renfrew 1999

The Tocharian split

Tocharian

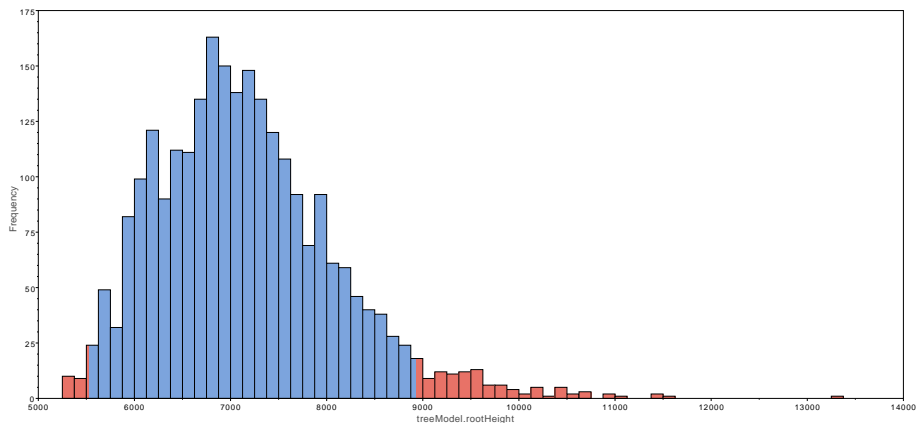
- Steppe theory (Anthony): migration from the steppe towards the Altai mountain ca. 3300–3000 BCE
- Anatolian theory (Renfrew 1999): eastward migration from the Balkans into the steppes
- archaeological evidence for intense cultural exchange between steppes and the Balkans by the end of the 4th millenium
- if identified with Tocharian split, this also gives a time constraint ca. 3300–3000 BCE

Replication of Bouckaert et al



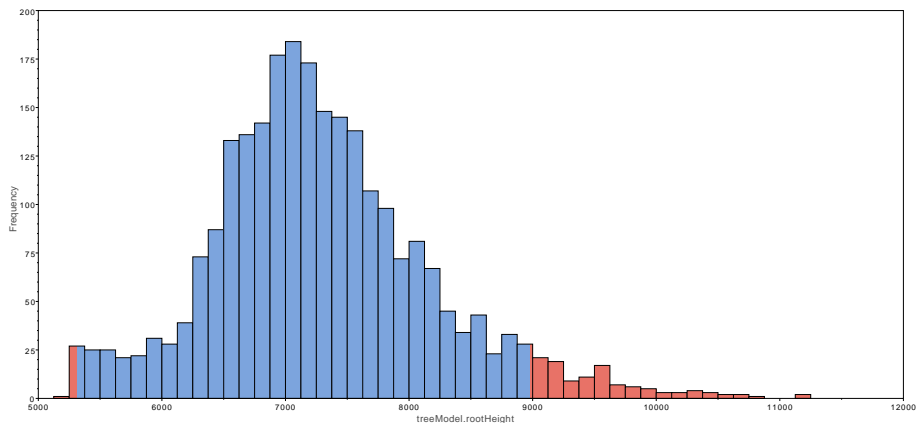
mean: 7685 BP; 95% HDI: 6053–9545

Adding Indo-Iranian constraint



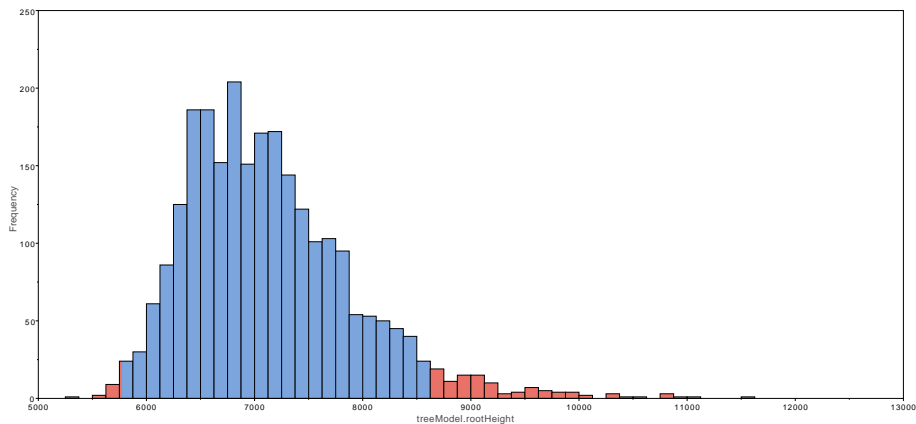
mean: 7162 BP; 95% HDI: 5525–8931

Adding Tocharian constraint



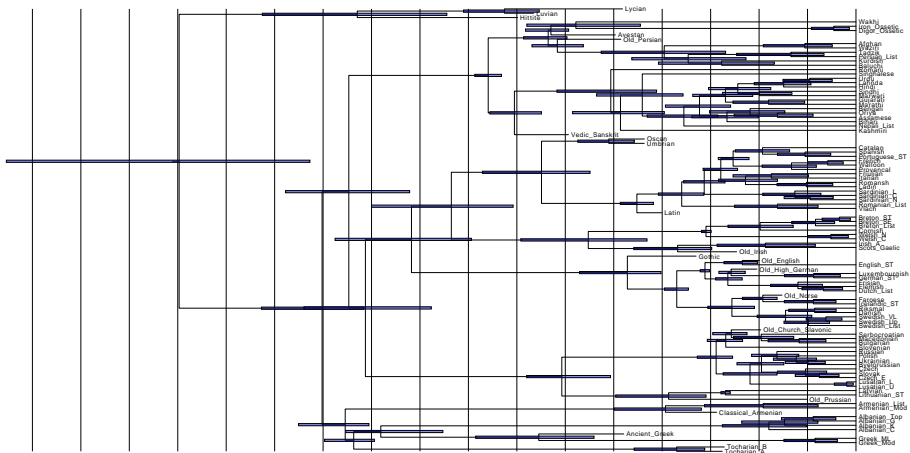
mean: 7295 BP; 95% HDI: 5316–8982

Adding both constraints

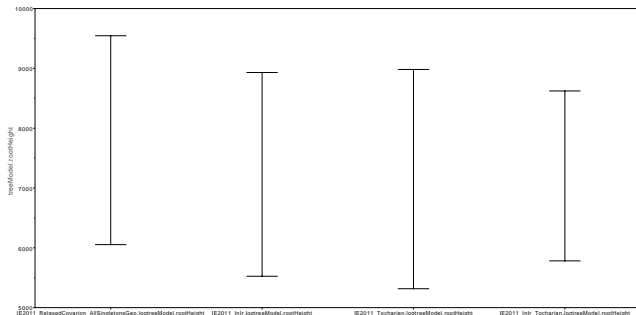


mean: 7139 BP; 95% HDI: 5782–8622

Adding both constraints



Summary



Bayes factor Steppe Theory vs. Anatolian Theory goes up from 1:1 to 4:1

And now for something completely different...

Borrowings in IELex

- general policy: loanwords are manually identified and marked as such
- usually excluded in phylogenetic analyses
- typical example: *mountain* (borrowed into English from French)
- sometimes decision is controversial:
 - Russian *sobaka* 'dog' is assigned to the same cognate class as German *Hund*, but there is a debate in the literature whether it is an Iranian or a Turkic loan (it is certainly not inherited)
- deep borrowings might not be recognizable as such

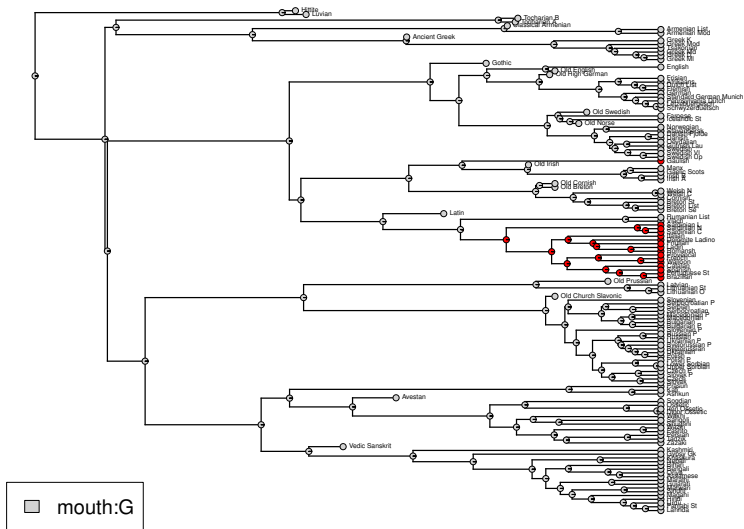
Impact on phylogenetic inference?

Semi-automatic loanword detection

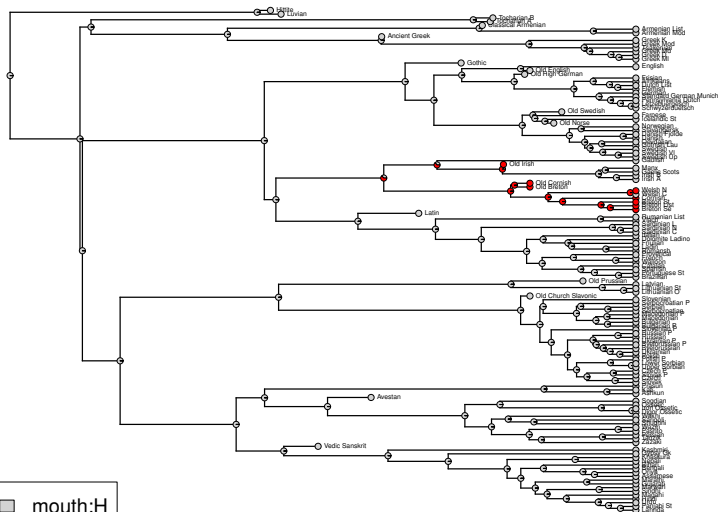
- In an ideal phylogeny, each cognate class should emerge exactly once (it might get lost several times).
- \Rightarrow Dollo-model of character evolution
- parallel changes $0 \rightarrow 1$ possibly indicative for borrowing
- There might be other reasons, such as
 - parallel semantic change
 - wrong phylogeny
 - incomplete sampling of synonyms
 - ...
- this talk:
 - Maximum likelihood character state reconstruction
 - manual identification of parallel changes

Semi-automatic loanword detection

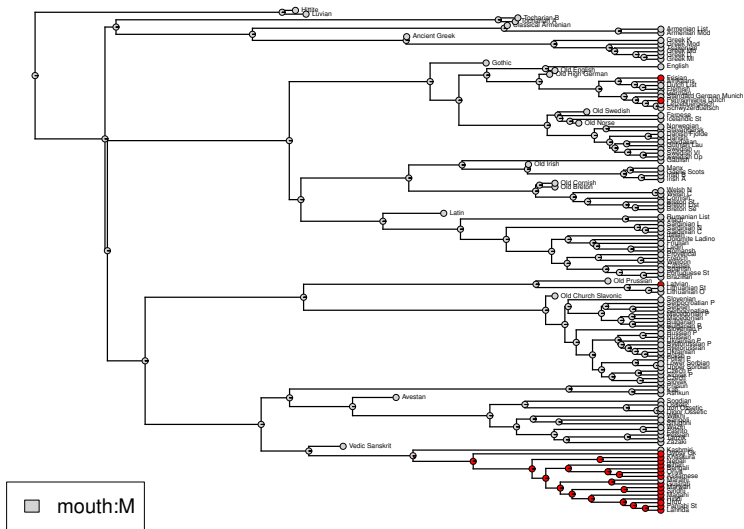
candidate for borrowing from Romance into Celtic



Semi-automatic loanword detection



Semi-automatic loanword detection



Semi-automatic loanword detection

- in total 109 out of 2,496 (non-trivial) characters were manually identified as probable loans
- Bayesian phylogenetic analysis, using the constraints from Bouckaert et al.,
 - with full and with “cleaned” data set
 - using all 150 varieties from IELex
 - using different software and tree prior (BayesPhylogeny, Yule tree prior)

