



Reconstructing language evolution using syntactical features

Robin Ryder, Jotun Hein & Thomas Mailund
Department of Statistics, University of Oxford, UK



Abstract

- Phylogenetic methods have been applied to historical linguistics previously, with some success.
- We attempt to reconstruct evolutionary trees of languages using syntactical features.
- We succeed in reconstructing most uncontroversial language families and genera. We test some more controversial groupings, and find no evidence for them.

Introduction

English	Latin	Greek	Sanskrit	Avestan	Tocharian A
father	pater	patér	pítár-	pitar-	pa:car
mother	ma:ter	má:ter	ma:tar-	ma:tar-	ma:car
foot	pedus	podós	pát, pá:dam	pad-	päts
salt	sa:l	háls	sal-ilá		sa:le
young	iuvenis		yúva	yavan	
red	rūbidos	ereuthos	róhita-	raoiðita-	rätr-ärkyant

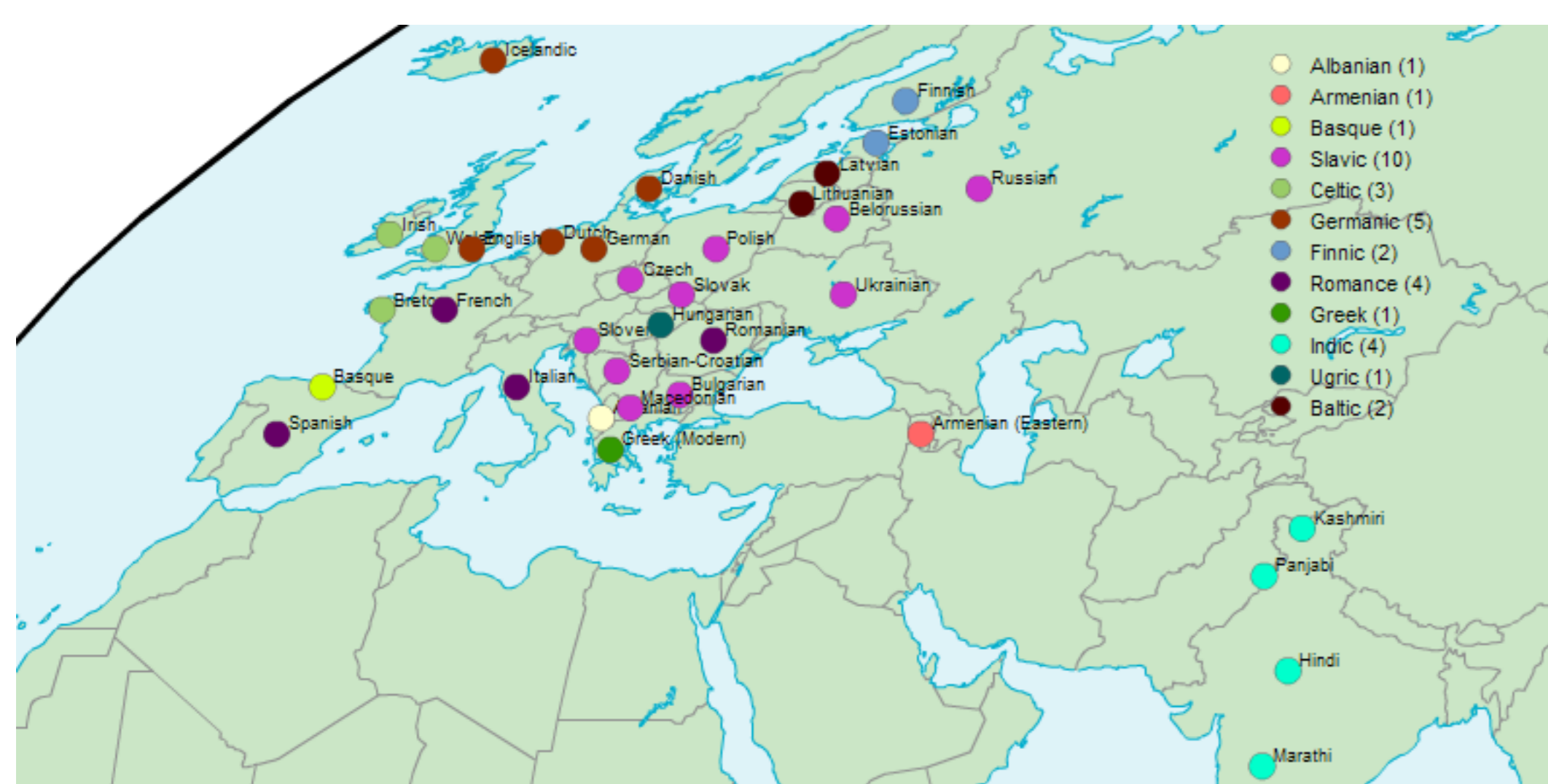
Numerous languages of Europe and India display similar roots in many words, indicating that they all evolved from one common ancestor. They are called the Indo-European languages.

There are similarities between language evolution and genetic evolution, so there have been several attempts to apply methods from computational biology to historical linguistics, usually by comparing lexical items.

Syntactical features (grammar) evolve in a similar way; we therefore wanted to know whether it was possible to reconstruct language history using exclusively these features.

Indo-European languages

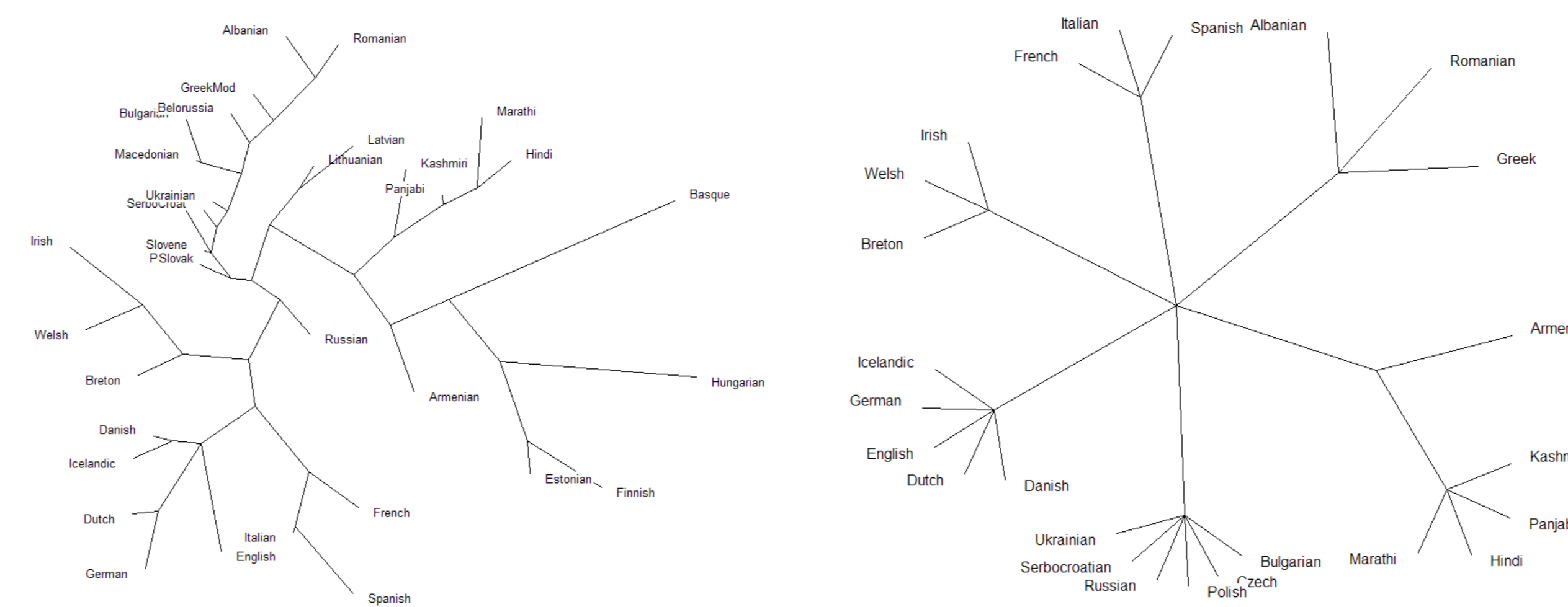
Indo-European languages have been well studied, so we used them to test the methods. There is hardly any controversy in the splitting of Indo-European languages into "genera".



This map shows the 35 languages included in the study of the Indo-European family, by genus.

Trees obtained

We compared two different methods of building evolutionary trees: maximum parsimony and a Bayesian analysis.



Left, tree obtained using a maximum parsimony method. Right, result given by Bayesian analysis (the displayed clades are those with posterior probability >.95).

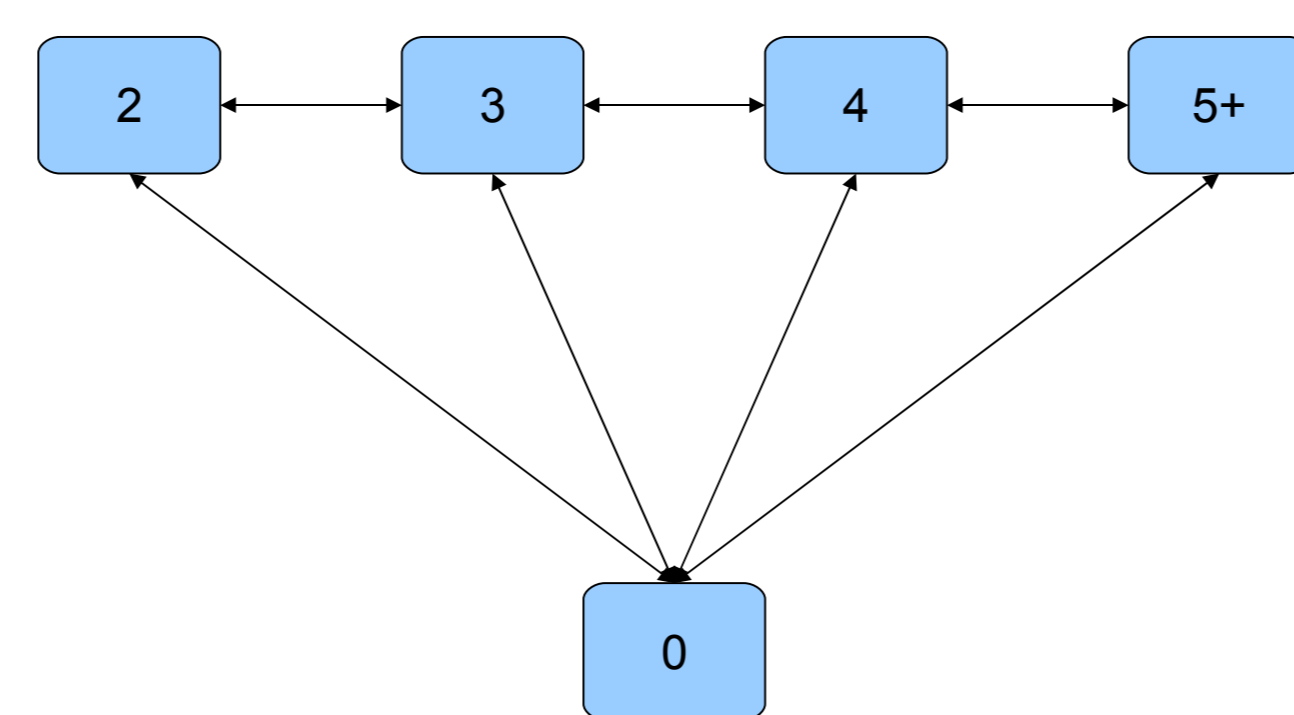
The tree on the right is almost exactly the one historical linguists have built using other methods (only Romanian is misplaced).

This leads us to believe that the same method could be applied to problems where the results are more controversial.

Model of grammar evolution

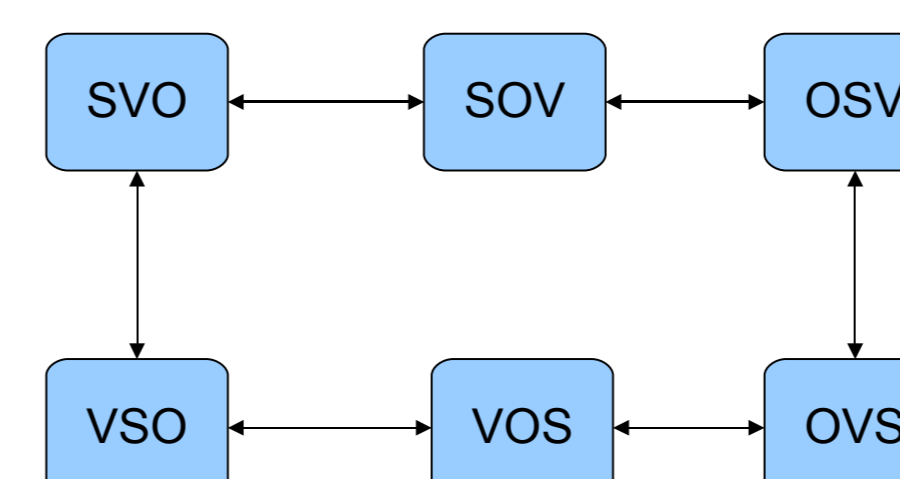
Before making any applications, we refined our data in two ways, which improved our results:

1. Using a different set of languages, we selected those features that seemed to be evolving the slowest, and therefore contain the most evolutionary information.
2. For many features, some traits seem closer to each other than others, so we introduced a model of evolution for those traits. Here are two typical models, for number of genders and basic word order:

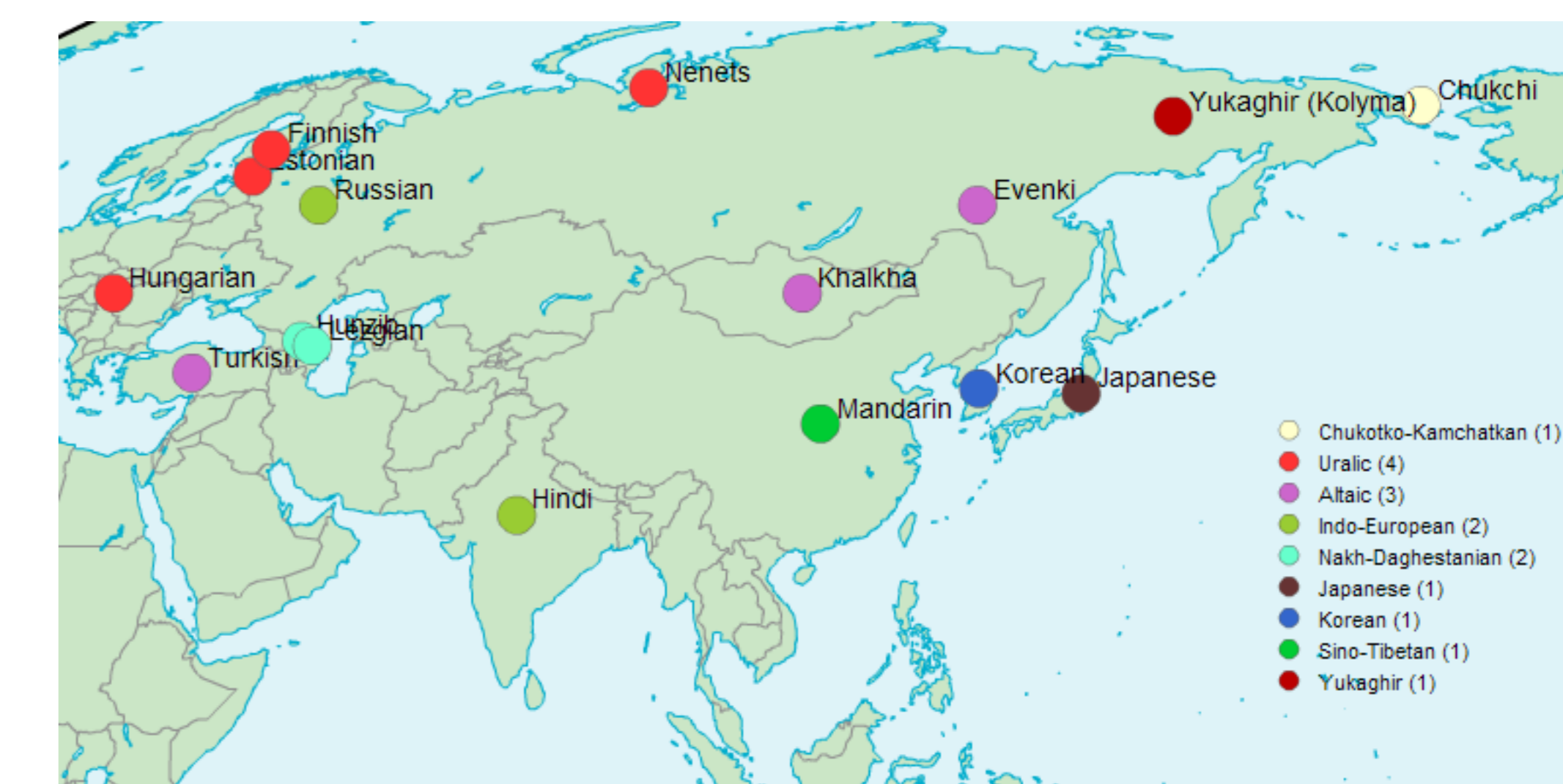


Model for the evolution of the number of genders

Model for the evolution of order of Subject, Object and Verb



Languages of Eurasia



Map of 16 languages of Eurasia. Some of the groupings shown are undisputed, but the Altaic grouping is controversial, so we wished to test it, along with other proposed families not shown here.

We find a clear split between the Indo-European, Uralic and Nakh-Daghestanian families on one hand, which are very well supported by our results, and tested groupings such as the Altaic or Uralo-Siberian families, for which we found no evidence at all in our results.

Possible Developments

This approach seems to confirm results that were generally already well accepted. It could be used to test other controversies in historical linguistics.

- However, several improvements should be made before that:
- Some grammatical features are correlated to others, so evolution of different features is not independent; this needs to be addressed.
 - More thought needs to be put into distinguishing resemblances due to a common ancestor, and those due to borrowing between languages.
 - The models of the evolution of syntactical traits are still very crude.

Phylogenetics could also provide historical linguistics with results that are not available with more traditional methods. The methods used allow us to date the ancestor, but there is not enough data to do so yet. Phonological and lexical data would probably be needed in addition to syntax.

References

The maps were made using the electronic version of the *World Atlas of Language Structure* (Dryer Hapselmath, Gil, & Comrie, 2005), which also provided most of the data. The data was analysed with Phylip (Felsenstein, 2005) and MrBayes (Huelsenbeck & Ronquist). The trees were made using TreeView (Page, 1996).