NLP Assignment 1, Spring 2024

Dr. Frederik Hartmann

Terms and Conditions

This is assignment 1 for the course *NLP*; the deadline for the assignment is 2/21/24, 11.59pm. The conditions for this assignment are as follows:

- The assignment has to contain **at least** 1,000 words for Master's students and 500 words for Bachelor's students, no abstracts and no bibliography are allowed. Assignment title, contents of tables, figure captions, model code, and section headers **do not** count towards the word count.
- Two file types need to be submitted to Canvas: the assignment itself and the code file(s) which you used in the analysis. **Do not** include programming code in your main assignment text.
- The file formats accepted are .rev, .txt, .py, .ipynb, .pdf, .docx. **Do not upload files in zipped folders!**
- The programming needs to be done in Python and RevBayes. Plotting software such as Tracer, FigTree, or IcyTree are allowed.
- Please name your files according to the following convention:
  - For **the documentation**:
    "DocumentationAssignment[AssignmentNumber]_[YourName].[FileEnding]"
  - For **the code files**:
    "CodeAssignment[AssignmentNumber]_[YourName]_[Filenumber].[FileEnding]"
- Everything pertaining to the documentation (i.e., everything that is not code) needs to go in a **single** *documentation* file.
- The code needs to be reproducible, i.e. it needs to run and produce the same output as reported in your documentation.
- Answer and discuss the assignment questions (see below) in **prose** directly, **no** bullet-point answers are allowed.

Failure to turn in the assignment by the end of the deadline and/or failure to adhere to the rules and conditions outlined above results in an F on this assignment. There will be **no extensions of the deadline**, **no acceptance of late submissions**, no and **no bonus assignments**. All files need to be submitted to Canvas only! The files that are submitted to Canvas at the time the deadline passes, are the ones that will be graded. You can revise and submit files as often as you want during the assignment availability period (i.e., *before* the deadline).

Make sure you leave enough time before the deadline to submit your files in case of technical issues. We are not responsible for software/connection issues on your end or related to Canvas. If you experience problems with Canvas specifically, please contact UNT IT support: https://teachingcommons.unt.edu/teaching-handbook/teaching-online/using-canvas

Assignment

This assignment consists of two parts, **both** of which have to be completed to receive full points.

Assignment part 1

On Canvas, you will find a dataset called *assignment_df.csv* which contains a binary dataset with 5 Romance languages plus Latin.

First, construct two distance matrices of the languages in the dataset, one with euclidean distance and the other with the Dollo model (dice distance). Afterwards, apply the UPGMA algorithm to both and plot the results. Interpret and describe the result in prose. Compare the results that both distance methods yield and discuss briefly why they might yield different or similar results by referencing the differences between the distance algorithms.

Next, construct a NeighborJoining tree or network from the dice distance matrix and plot the results. Here too, interpret the results and compare them with the results of the UPGMA results of the same distance matrix by referencing the difference between UPGMA and NeighborJoining algorithms.

Assignment part 2

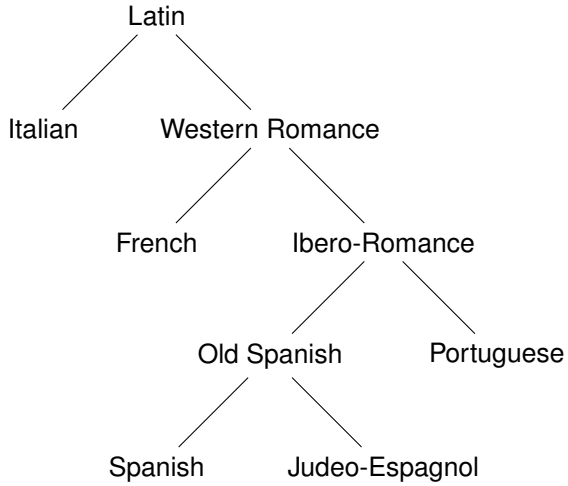
On Canvas, you will find a nexus-format dataset called *assignmentdata.nex* which contains a binary dataset with 5 Romance languages plus Latin.

Write a RevBayes phylogenetic model that fulfills the following criteria:

1. Is an actual-time calibrated phylogenetic model that infers the root age
2. Has Latin as a fossil with a uniform time uncertainty interval between 500 BC and 1 AD
3. Assumes that Latin is equal to the root.
   There are multiple ways of doing this, so there is no *single* right way of doing this. **Hint**: try setting the fossil age of Latin equal to the root age.
4. Assumes that Judeo-Espagnol and Spanish split somewhere between 1500 AD and today.
5. Includes:
   (a) gamma-distributed site-rate variation
   (b) variable branch rates
   (c) root frequency estimation

In the text, describe the model and justify your modelling and prior choices. Please interpret further the posterior estimates of the site rate parameter (α), the root age parameter, and the Q-matrix parameters.

In a last step, plot the MCC or consensus tree and interpret the tree topology by comparing it to the UPGMA tree obtained from the dice distance matrix in Assignment part 1 (above) and the Romance family tree obtained through non-computational diachronic methods (below). Discuss the differences and similarities of all three trees and briefly discuss why these differences/similarities might arise by referencing the different methods with which the trees were constructed.



Traditional Romance family tree