# Portfolio Exercise 1

25 April 2018

Quantitative Methods in Historical Linguistics

Henri Kauhanen

This exercise is worth a maximum of 5 points.

Important: Make sure you include all R code you used to figure out your answer!

1. First, download the file `phoible_inventories.csv` from the course page on ILIAS (it's in a folder named 'Data').
2. Import this into R using the `read.csv` function, storing it in a variable named `phoible`.
3. Type `head(phoible)` to get an idea of the structure of the dataframe. (This command shows you the first five rows of the dataframe.)
4. Try `View(phoible)`, too.
5. Now find out the average size of consonant inventories. (Hint: use R's `mean` function on the relevant column of the dataframe.)
6. Do the same for the average size of vowel and tone inventories.
7. Read the short document 'Mean, Median and Mode' (on ILIAS).
8. Now repeat points 5–6 above using R's `median` function instead of `mean`.
9. You should find that the median size of tone inventories is 0. What does this mean?
10. What is the size of the consonant, vowel and tone inventories of (a) English and (b) Vietnamese?

# Portfolio Exercise 2

9 May 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 15 points.

The table on the following page gives 35 cognates in five Quechuan languages. The cognates are given in IPA transcription; '—' means 'no data available'.

1. Use the comparative method to reconstruct as many phonemes (both vowels and consonants) of Proto-Quechuan as you can based on these data. Your answer should contain also the sound correspondences you used for the reconstructions, as well as a justification for each reconstruction (majority wins / directionality / something else?). The sound correspondences may, if needed, refer to phonological environments (e.g. #p, V#, VbR, see lecture slides).
2. Next, find the most likely subgrouping (tree) for these five languages based on the data in the table, and draw the family tree. Please also include in your answer a justification: why did you arrive at this particular subgrouping?

TABLE (from Campbell, L. 2013. *Historical linguistics: an introduction.* 3rd edn. Edinburgh: Edinburgh University Press).

| | **Ancash** | **Junín** | **Cajamarca** | **Amazonas** | **Ayacucho** | **gloss** |
|---|---|---|---|---|---|---|
| 1 | paka | paka | paka | paka | paka | ‘begin’ |
| 2 | apa | apa | apa | apa | apa | ‘wash’ |
| 3 | rapra | lapla | rapra | rapra | rapra | ‘leaf’ |
| 4 | pampa | pampa | pamba | pamba | pampa | ‘plains’ |
| 5 | tapu | tapu | tapu | tapu | tapu | ‘ask’ |
| 6 | wata | wata | wata | wata | wata | ‘tie’ |
| 7 | utka | utka | utka | utka | utka | ‘cotton’ |
| 8 | inti | inti | indi | indi | inti | ‘sun’ |
| 9 | kimsa | kimsa | kimsa | kimsa | kimsa | ‘three’ |
| 10 | puka | puka | — | puka | puka | ‘red’ |
| 11 | haksa | saksa | saksa | saxsa | saksa | ‘be full’ |
| 12 | kuŋka | kuŋka | kuŋga | kuŋga | kuŋka | ‘neck’ |
| 13 | qam | am | qam | kam | χam | ‘you’ |
| 14 | qoha | usa | qosa | kusa | χosa | ‘husband’ |
| 15 | waɢa | waʔa | waɢa | waka | waχa | ‘cry’ |
| 16 | hoχta | suʔta | soχta | sukta | soχta | ‘six’ |
| 17 | tsaki | tʃaki | tʃaki | tʃaki | tʃaki | ‘dry’ |
| 18 | mutsa | mutʃa | mutʃa | mutʃa | mutʃa | ‘kiss’ |
| 19 | mantsa | mantʃa | mantʃa | mantʃa | mantʃa | ‘fear’ |
| 20 | putska | putʃka | putʃka | putʃka | putʃka | ‘thread’ |
| 21 | e:tsa | aitʃa | aitʃa | e:tʃa | aitʃa | ‘meat’ |
| 22 | utʃpa | utʃpa | utʃpa | utʃpa | utʃpa | ‘ashes’ |
| 23 | kitʃki | kitʃki | kitʃki | kitʃki | kitʃki | ‘narrow’ |
| 24 | haru | salu | saru | saru | saru | ‘step’ |
| 25 | hara | sala | sara | sara | sara | ‘corn’ |
| 26 | qaha | asa | qasa | kasa | χasa | ‘ice’ |
| 27 | iʃke: | iʃkai | iʃkai | iʃke: | iskai | ‘two’ |
| 28 | waʃa | waʃa | waʃa | waʃa | wasa | ‘behind’ |
| 29 | hatuŋ | hatuŋ | atuŋ | atuŋ | hatuŋ | ‘big’ |
| 30 | hutsa | hutʃa | utʃa | utʃa | hutʃa | ‘fault’ |
| 31 | humpi | humpi | — | umbi | humpi | ‘sweat’ |
| 32 | laki | ʎaki | ʒaki | dʒaki | ʎaki | ‘pain’ |
| 33 | kila | kiʎa | kiʒa | kidʒa | kiʎa | ‘moon’ |
| 34 | alba | aʎpa | aʃpa | adʒpa | aʎpa | ‘land’ |
| 35 | ailu | aiʎu | aiʒu | e:dʒu | aiʎu | ‘family’ |

# Portfolio Exercise 3

16 May 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 10 points.

Important: Make sure you include all R code you used to figure out your answer!

1. First, make sure you have the R packages needed to complete this exercise:

   ```
   library(rdist)
   library(phangorn)
   ```

   If not, install them as usual:

   ```
   install.packages(c("rdist", "phangorn"))
   ```

2. Download the dataset `wals_IE.csv` from the course's ILIAS page (it's in the folder 'Data').
3. Import this into R as follows, storing it in a variable named `wals` (for example):

   ```
   wals <- read.csv("wals_IE.csv", row.names=1)
   ```

   (Here, `row.names=1` means that the row names of the resulting dataframe will be taken from the first column of the CSV file.)
4. Type `wals` to see the data. You should see a feature matrix, with rows corresponding to 17 languages and the columns corresponding to 14 features (named `X129A` and so on).[1]
5. Turn this feature matrix into a distance matrix.
6. Apply UPGMA to the distance matrix.
7. Plot the resulting tree.
8. You should end up with a rooted tree with 17 leaves, all of them Indo-European languages. Next, do some research (on the internet, in the library...) on the Indo-European family. Compare your tree to trees of Indo-European published elsewhere, and comment on the similarities and differences. Do the trees look similar? What differences are there? Which languages are in the "wrong place" in your tree, in your opinion?

[1] The features are taken from the World Atlas of Language Structures (WALS) database. See http://wals.info.

# Portfolio Exercise 4

23 May 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 20 points.

## Part I

1. Intuitively, linguistic features which change too often are not useful in phylogenetic analysis, since they carry a high risk of homoplasy and back-mutation. Such features are known as *unstable* features. The opposite is *stable* features, which are much less likely to change within a language family. To find out more about unstable and stable features and how featural stability may be measured, read this article: https://www.researchgate.net/profile/Soren_Wichmann/publication/292655546_Diachronic_stability_and_typology/links/570d38c908aed31341cf7454.pdf.
2. In Portfolio Exercise 3, you used UPGMA to generate a tree of Indo-European based on 14 typological features. The tree was not perfect. Based on your reading of the above article, which of the 14 features are to blame? (To read more about the features, browse to http://wals.info. Feature X129A in the dataset corresponds to WALS feature 129A, and so on.)

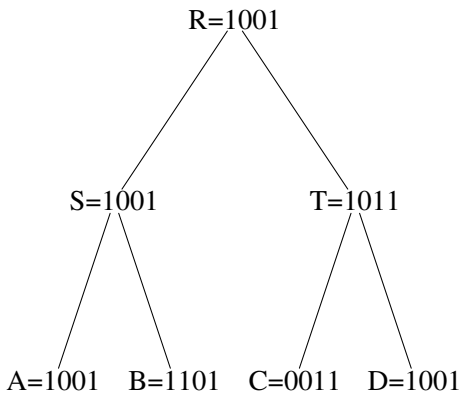
## Part II

3. Suppose we have four languages A, B, C and D, and four binary features F1, F2, F3 and F4, with the following feature matrix:

| | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 1 |
| B | 1 | 1 | 0 | 1 |
| C | 0 | 0 | 1 | 1 |
| D | 1 | 0 | 0 | 1 |

   Use UPGMA (by hand, not by using R) to infer the topological phylogenetic tree. Show all steps of the calculation.

4. Now suppose the true phylogenetic tree of these languages is the following:




where R, S and T are ancestral languages with the feature vectors 1001, 1001 and 1011, respectively. What property of this historical evolution leads the distance-based UPGMA algorithm astray?

# Part III

5. How many different binary-branching trees are there of three leaves A, B and C? Draw all of them.
6. How many different binary-branching trees are there of four leaves A, B, C and D? Draw all of them.

# Portfolio Exercise 5

23 May 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 10 points.

Important: Make sure you include all R code you used to figure out your answer, as well as the lexical and distance matrices.

1. Go back to Portfolio Exercise 2 on Quechua. Import the data in the cognate table to R and make a lexical matrix. (Note: it is best to replace the difficult-to-type IPA characters (e.g. ʃ) with something else. You can use capital letters for example, since R is case-sensitive. E.g. tʃaki becomes tSaki.)
2. Turn the lexical matrix into a Levenshtein distance matrix.
3. Use UPGMA to infer a tree from the distance matrix, and plot the tree.
4. Compare the tree to your subgrouping in Portfolio Exercise 2, and comment on any possible differences between the trees. If the trees are different, which tree do you believe is the better representation of reality?

# Portfolio Exercise 6

6 June 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 10 points.

1. Write an R function that takes three arguments, $t$, $s$ and $k$, and gives the value of the logistic function
$$f(t) = \frac{1}{1 + \exp(s \times (k - t))}$$
as output.
2. Write an R function that takes five arguments, $t$, $s$, $k$, $A$ and $B$, and gives the value of the modified logistic function
$$f_{\text{mod}}(t) = A + \frac{B}{1 + \exp(s \times (k - t))}$$
as output.
3. What do the parameters $A$ and $B$ do?
4. What values of $A$ and $B$ are possible parameter values if $f_{\text{mod}}(t)$ is to be interpreted as a relative frequency?
5. Read this webpage:
`http://www.simonqueenborough.info/R/stats-advanced/functions`
to learn how to specify default argument values when writing R functions. Then write an R function that takes five arguments, $t$, $s$, $k$, $A$ and $B$ as input and gives $f_{\text{mod}}(t)$ as output, and has the values $A = 0$ and $B = 1$ as default for the arguments $A$ and $B$ (but no default values for the arguments $t$, $s$ and $k$).

# Portfolio Exercise 7

27 June 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 10 points.

Important: Your answer should include all the R code you used, plus any plot(s) you need to illustrate your answer to point 3 below.

1. Working on the dataset `ellegard_full.csv`, fit logistic models to each of the contexts: affirmative questions, negative questions and negative declaratives. Then use the `cretest` function (in `cretest.R`) to test whether this is a CRE.
2. Repeat the above, but this time only use the time periods from 1400–1425 to 1550–1575 (i.e. the first 7 rows). Do we have a CRE now?
3. Plot the relative frequencies from `ellegard_full.csv` against time period midpoints. What happens after the period 1550–1575? Why is it a good idea to do what we did in point 2 above, i.e. to exclude time periods after 1550–1575?

# Portfolio Exercise 8

4 July 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 10 points.

1. Working on the toy model ("A first model") introduced in the lecture slides "Dynamical systems 1" (4 July 2018), calculate the values $x_1, x_2, \ldots, x_{10}$ and $y_1, y_2, \ldots, y_{10}$ when $x_0 = 100$, $y_0 = 200$ and $a = 0.5$.
2. Use R to produce a scatterplot or lineplot with $t$ on the horizontal axis and $x_t$ and $y_t$ on the vertical axis (both $x_t$ and $y_t$ in the same plot).
3. If you had to make a guess, what would you say the values of $x_{100000}$ and $y_{100000}$ are?

# Portfolio Exercise 9

11 July 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 20 points.

# Part I

1. Write a for-loop that iterates the second model introduced in the slides "Dynamical systems 2" (11 July 2018), and wrap this loop up in a function.
2. Use your function to iterate the model with parameter values $a = 0.1$ and $b = 0.2$ from the following two initial conditions:
   (a) $x_0 = 10$, $y_0 = 990$
   (b) $x_0 = 990$, $y_0 = 10$

   Iterate the model for as many iterations as is necessary to see the long-term behaviour of the model.
3. Then plot $x_t$ and $y_t$ against $t$ (i.e. $x$ and $y$ on the vertical axis, $t$ on horizontal axis)
4. Find the fixed point of the system.
5. Now suppose $a = 0.01$ and $b = 0.02$. What is the fixed point?
6. Iterate and plot the model with $a = 0.01$ and $b = 0.02$. What is the difference, compared to when $a = 0.1$ and $b = 0.2$?

# Part II

Let's continue working with the "second model". However, suppose now that our language community of $X$-speakers and $Y$-speakers is geographically adjacent to another community, all of whom are $X$-speakers. Let's assume that, in addition to the already familiar processes $X \to Y$ (at rate $b$) and $Y \to X$ (at rate $a$) occurring in the first community, there is language contact between the two communities. And let's assume that this contact means that, during one time step, a proportion $c$ of the $X$-speakers in the first community who would normally become $Y$-speakers over that time step actually stay $X$-speakers (because of the influence from the second, entirely $X$-speaking community).

1. Write the difference equations for this kind of model. (Note: you do not need an equation for the second community. We are only assuming that the second community affects the speakers of the first community.)
2. Figure out the fixed point (there is only one) of this model.

# Portfolio Exercise 10

18 July 2018

Quantitative Methods in Historical Linguistics

Dr. Henri Kauhanen

This exercise is worth a maximum of 15 points.

1. Write an R function that iterates the third model from the lecture slides "Dynamical systems 3" (18.7.2018).
2. Use your function to iterate the model for 100 time steps, using the initial values $x_0 = 100$ and $y_0 = 500$, (a) with $d = 0.1$, and (b) with $d = -0.01$. Plot the curves.
3. Iterate the model with $d = 3$, $x_0 = 10$ and $y_0 = 590$ for 100 time steps. Plot the result. What do you see?
4. Does the parameter value $d = 3$ have a meaningful real-world interpretation? Between what numbers must $d$ be in order to make sense? (Hint: Remember that $d = a - b$, and think about what $a$ and $b$ meant in the original model.)
5. Make state space plots of the second and third models for some non-zero parameter values $a$, $b$ and $d$ of your own choosing. How do the plots differ from each other? Why do they differ in this way?
6. Use the cobwebbing method to show that the fixed point of the second model (assuming $a, b \neq 0$) is attracting.