This exercise counts 25% toward your final grade. There are multiple exercises to chose from (2A or 2B; within 2A there are four options). Pick the one that best suits your interests. Either can be further developed in the final report if you wish to.

Exercise 2A: Character-based language models

There are different ways you can approach this exercise. You should focus on one only. The general task is the same for everyone.

General task Exercise 2A:

Train a character-level language model to generate names (persons, companies, animals, etc) in your language of choice[1]. Assess your model's quality. Feel free to use A. Karpathy's code as a template.

Specific task 2A-I: Multiple models; multiple languages

Pick a second language. Build a second character-level model for this language and assess its quality. Then assess your first model on the task of generating names for the second language, and vice-versa. How good is the transfer from the first model to the second? How about the other way around? Is this what you expected? Why (not)?

Specific task 2B-II: One model; multiple periods of time

Deploy your model on a different dataset of the same kind but from a different time period. For instance, if you were generating modern Dutch names, also evaluate how well the model fares when generating Dutch names from the 1900s. Pick whichever time period you like. You are free to explore how the model fares across more than two time periods but this is not required. Assess the model's quality in this diachronic transfer task. Is this what you expected? Why (not)?

Specific task 2C-III: One model; multiple evaluations

Evaluate your model's quality in at least three different ways. Discuss what (dis)advantages each method has. Discuss the model's overall quality in view of your findings.

Specific task 2D-IV: Non-latin script

If your data comes from a script that is not Latin-based then –apart from assessing its quality– describe the changes you had to make to get it to work. Focus both on linguistic features of the language you chose and its script in your discussion of changes and quality. You can additionally also do any of the (sub)tasks from I-IV but this should not be your focus.

The report should be a PDF document; no longer than 2 pages. It should include the sections described below. The 2-page limit does not include references (optional); supplementary material (optional); and the list of contributions (only if working in groups).

Introduction (5%)

Briefly explain the task in your own words.

[1] If you pick English then it has to be something other than person names (i.e., something different than what was already done in the tutorial)

Material and methods (25%)

Describe your approach to as well as the data you will be using (where did you get it? how did you process it?)

Results (40%)

Your main results. Discuss limitations of your data and methods.

Code (30%: 20% replicability/10% clarity)

Make your code publicly available (hosted on, e.g., OSF or github). Remember to extensively comment it. Mention the dependencies that need to be fulfilled to run the code.

List of contributions (unlimited space / only if working in groups)

Who did what in your group. You can use the CRediT system or a variant thereof.

Exercise 2B: Byte pair encoding (BPE) and morphology

Study to which extent BPE aligns with the traditional notion of a morpheme for a language of your choice by deploying it on a corpus. You are free to decide how you go about this in terms of BPE-implementations, hyper parameters, corpora, and means of comparison.

The report should be a PDF document; no longer than 2 pages. It should include the sections described below. The 2-page limit does not include references (optional); supplementary material (optional); and the list of contributions (only if working in groups).

Introduction (5%)

Briefly explain the task in your own words.

Material and methods (25%)

Describe the data you will be using (where did you get it? how did you process it?) and how you will compare the compressed outcomes to a more traditional morphological analysis of the data.

Results (40%)

Your main results. Discuss limitations of your data and methods.

Code (30%: 20% replicability/10% clarity)

Make your code publicly available (hosted on, e.g., OSF or github). Remember to extensively comment it. Mention the dependencies that need to be fulfilled to run the code.

List of contributions (unlimited space / only if working in groups)

Who did what in your group. You can use the CRediT system or a variant thereof.