

# Assignment 1

Deadline: Friday, 29th of September, 11:59pm

## Notes

- You can work in groups of 2 or 3 students. Please state clearly in your accompanying document who was part of your team (full names and student ID numbers)
- Data sets for the assignment can be found on  
<https://staff.fnwi.uva.nl/m.fadaee/ALT/file.en>  
<https://staff.fnwi.uva.nl/m.fadaee/ALT/file.de>  
<https://staff.fnwi.uva.nl/m.fadaee/ALT/file.aligned>
- The bitext contains 50,000 lines of tokenized and lowercased sentences (`file.en` and `file.de`) and a (refined) word alignment `file.aligned`.

## Translation Model

Your task is to build a translation model. You have to implement and compute the following for both data sets (clean and web):

### 1. Phrase extraction algorithm

Output: program and file with extracted phrases. The file should be of the form:

$f \parallel e \parallel freq(f) \ freq(e) \ freq(f, e)$

### 2. Phrase translation probabilities

Output: program and file containing probabilities  $p(f|e)$  and  $p(e|f)$ .

The file should be of the form:

$f \parallel e \parallel p(f|e) \ p(e|f)$

### 3. Lexical translation probabilities (KMO approach)

Output: program and file of the form:

$f \parallel e \parallel p(f|e) \ p(e|f) \ l(f|e) \ l(e|f)$

### 4. The resulting files of (1-3) can be combined into one single file of the form:

$f \parallel e \parallel p(f|e) \ p(e|f) \ l(f|e) \ l(e|f) \parallel freq(f) \ freq(e) \ freq(f, e)$

## Submission

- Write a short report (in pdf) and include:
  1. Description of how to run your programs.
  2. Discussion (up to two pages) of a few phrase translation examples where the translation probabilities are clearly counterintuitive. Also discuss how you could address some of these issues.
- Submit the code (code only, not the files!) and the report together as a gzipped tarball via blackboard.
- Place the files on a file sharing website of your choice and submit the path. Make sure the path is accessible and readable by others.