Graded assignment 3:
Dispersion

1. Pseudocode assignment

In the <files> folder of this course, there is a folder called <ICEGB_sampled>, which contains a reduced version of the 500 files of the British Component of the International Corpus of English (ICE-GB). These files look like this (i.e. each word in the corpus is surrounded by curly brackets):

```
 DISMK,FRM {OK}
  NPHD,N(prop,sing) {Adam}
 DISMK,INTERJEC {uhm}
 PAUSE,PAUSE(short) {<,>}
  NPHD,PRON(inter) {what}
 INTOP,AUX(do,past) {did}
  NPHD,PRON(pers) {you}
  MVB,V(cxtr,infin) {see}
  P,PREP(ge) {as}
    MVB,V(intr,ingp) {missing}
[…]
```

Our goal is to compute the dispersion statistic *DP* for a handful of words. Write pseudocode and note down function names to prepare you to develop code that will, ultimately,

1. load each of the 500 files of the ICE-GB
   - compute its size in words
   - prepare to be able to find out quickly the frequency of each word for which we want the dispersion;
2. compute the relative size of each corpus file out of the whole corpus (essentially its percentage of the whole corpus);
3. compute the relative frequency of *the* in each corpus file out of the whole corpus (essentially its percentage of the whole corpus);
4. compute *DP* from those vectors.

Submit this to Albert at <aventayolboada@ucsb.edu> by 28 Feb 2023, 08:00 PST.

