Data:

Pick a dataset from UniversalDependencies. For example Finnish-TDT for Finnish and English-EWT for English are good ones, but you can pick any other language of interest. You can wget these files directly from github, eg:  <https://github.com/UniversalDependencies/UD_Finnish-TDT/raw/master/fi_tdt-ud-train.conllu>

1. Familiarize yourself with the conllu format <https://universaldependencies.org/format.html>
2. Then write a little program which goes over this file, line per line, and gathers word statistics as follows:
   * For Finnish: They always say a Finnish verb has 10,000 different forms. How many of these do you see in reality, i.e. in this dataset? Ideally, you'd list them in a descending order of frequency. I leave the details to you. By verb form I mean the "morphological features" column.
   * For English: They always say you can *verb any noun* and *noun any verb*. How many cases do you see where this happens, i.e. for how many words do you find evidence for the same word being used as a noun and also as a verb in the data?
   * For any language: how big a proportion of the data (in terms of percentage of running words) is covered by words which only have one possible analysis (lemma, POS tag, features). I.e. how ambiguous that language is?

Upload your results as a .ipynb notebook, preferably, or a .py file with its output if needs be. :) If you fail to solve the whole thing, then upload what you have. An attempt counts!