What does a dataset need?

* Vocabulary (list of strings)
  + Source (biblical hebrew)
  + Target (modern heberew)
  + Include these: ['<pad>', '<unk>', '<s>', '</s>',
  + Each word gets an id. Keep those in a dictionary mapping word to id.
* Sentences (list of lists of strings (words)).
  + Source
  + Target
  + The same number of sentences in both
  + Take some out of here for validation.
* Dataset class
  + \_\_getitem\_\_
    - Returns a sentence that starts with SOS, ends with EOS, and pads all the way to MAX\_SENTENCE\_LENGTH.
  + Throw this into a dataloader later.
* Notes:
  + Removed \u2009, \u200f[1] -> both unicode for space
  + Removed words of length less than two
  + Removed possible psukim (כ״ח, י״ד…)
  + Kept square symbols: since Google succeeded with differentiating them, we figured the model would too.