+ It's just a sofesticated predection model.
+ we Transform our corpus into lists of Token.

So for one line of Text:

$[4, 2, 66, 70] \Rightarrow [4, 2]$

$\Rightarrow [4, 2, 66]$

$\Rightarrow [4, 2, 66, 70]$

+ we then pad the above sequences.

We now Take everything but the end
of The sequence on Y

1. $XS \Rightarrow$ All but last
2. labels $\Rightarrow$ last.

+ One hot encoding $\Rightarrow$ (labels, num-classes)
  Y

Total words.

So g will be an array with The length
of The whole corpus of words.

+ For less memory, character based
Prediction.