Lesson5:
Generative Models for Text on the Web
Unit4:
Increasing the number of model parameters

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

WeST

People and Knowledge Networks

Completing this unit you should

- See that one can always increase the model parameters

- Know that increasing model parameters often yields a more accurate model

- Be aware of the bigram and mixed models as examples for our generative processes

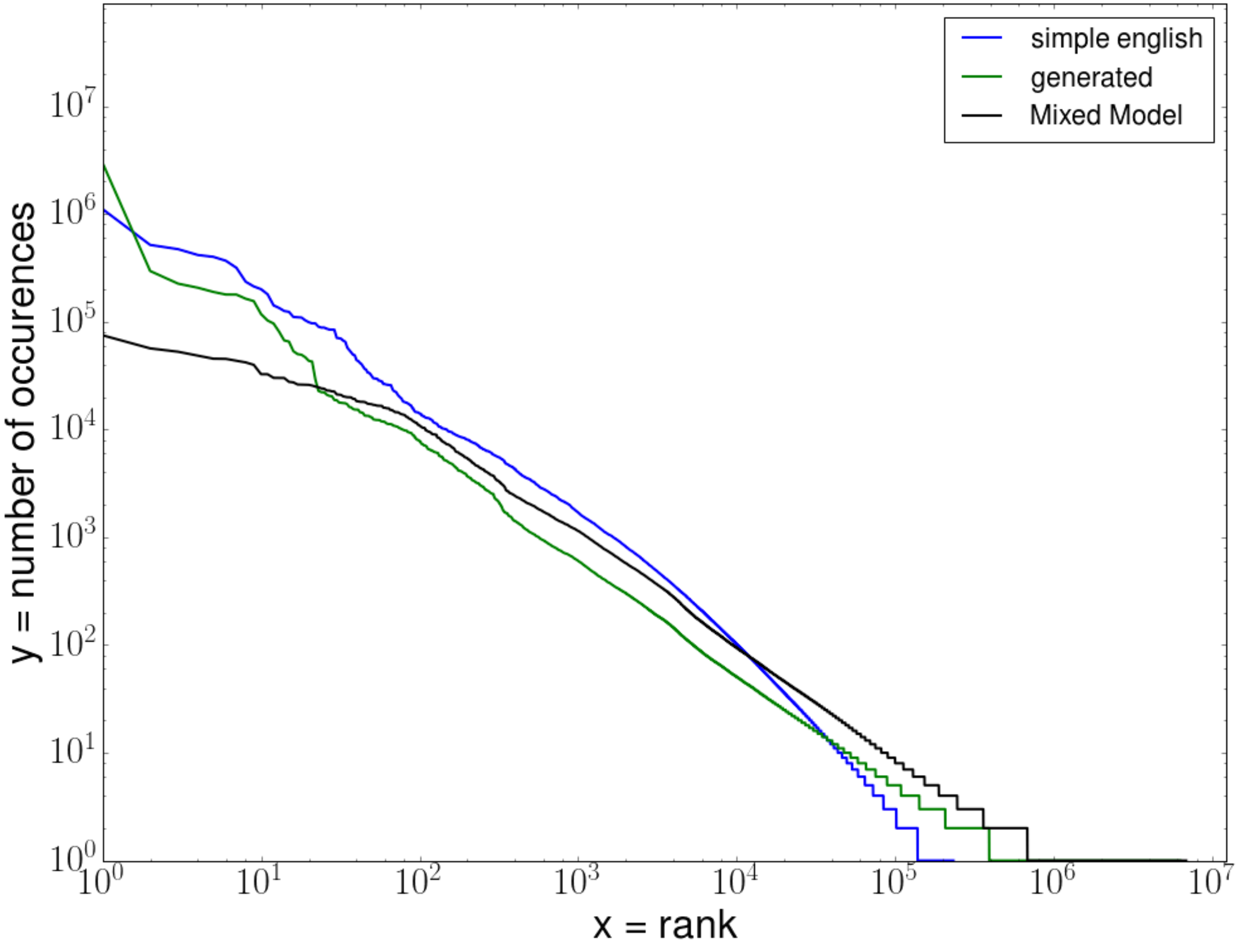
What happens if we try to encode the length into our model?

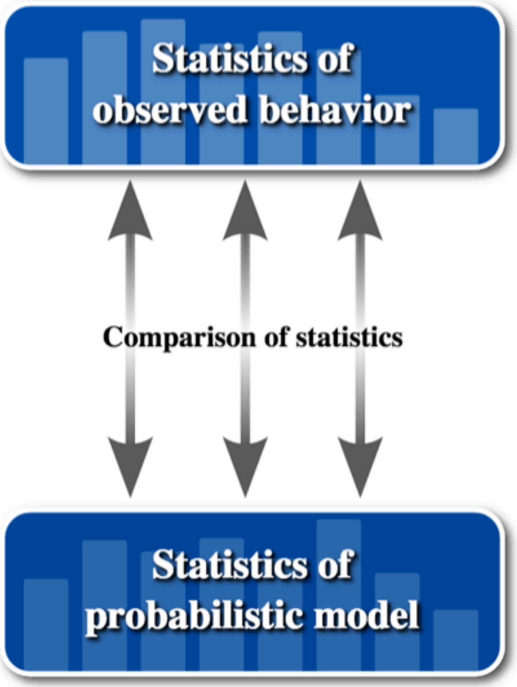
- Make a 2 step process

- First step learn the word length distribution
  - Randomly select a word length "n" for the next word that should be generated
- Learn the unigram distribution (without) space
  - Draw "n" characters from the unigram distribution
- We call this model the "mixed model"

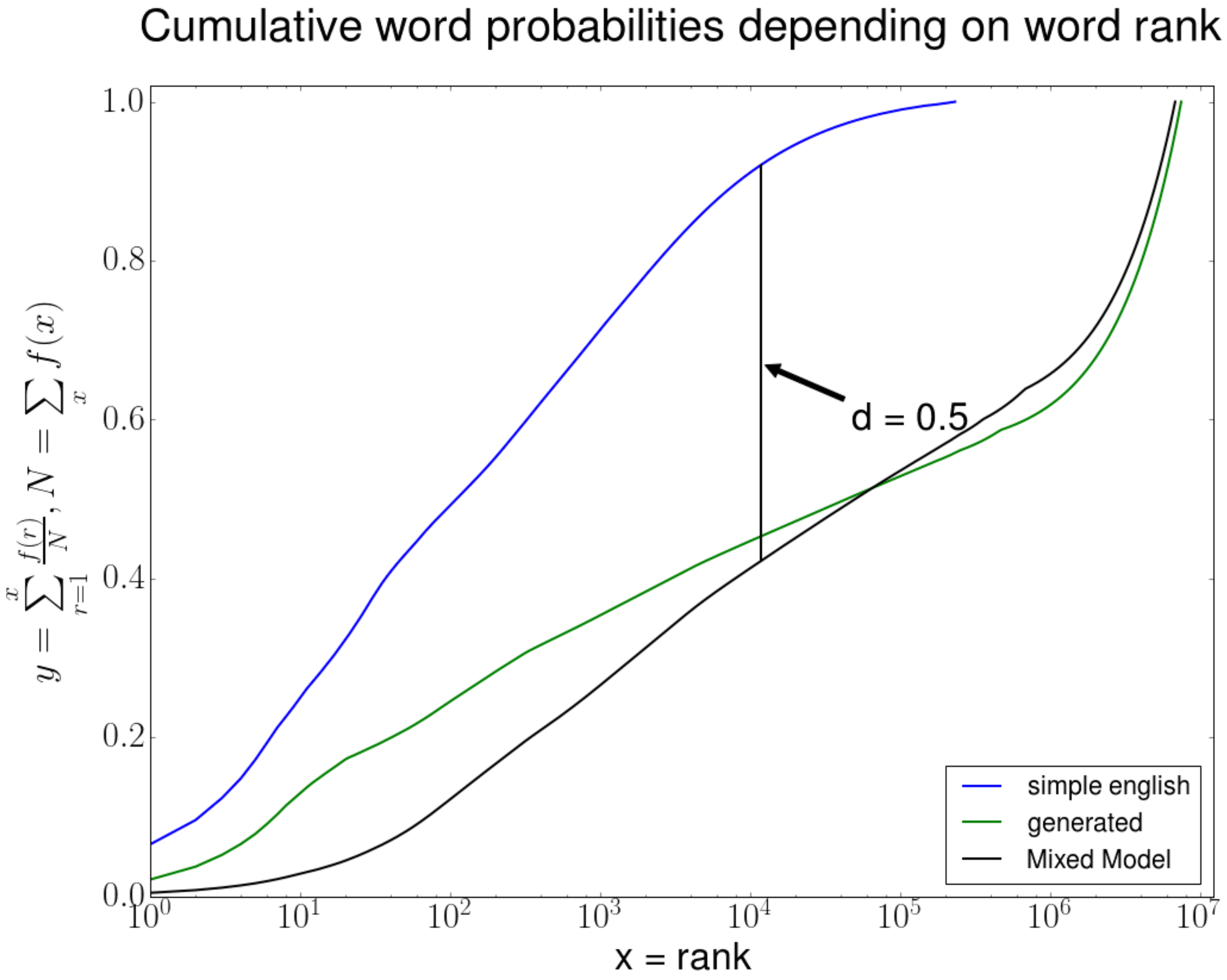
Plotting the Zipf distribution looks worse for the mixed model

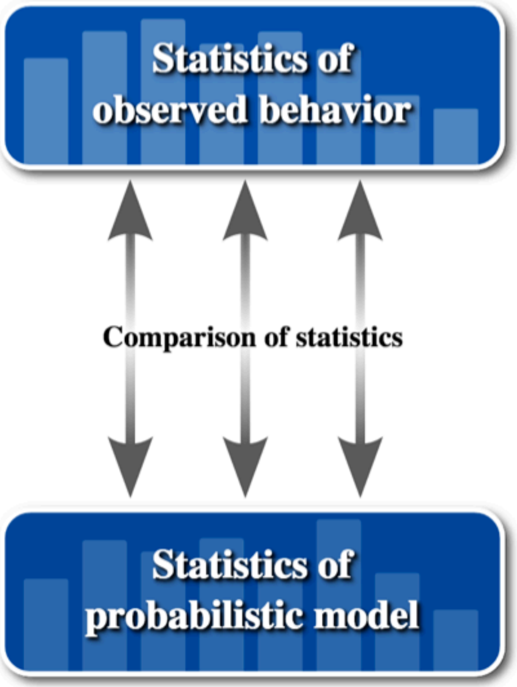
Word frequencies depending on word rank on (Simple) Englisch Wikipedia




Statistics of observed behavior
Comparison of statistics
Statistics of probabilistic model

Also cumulative plot verifies this





Statistics of observed behavior
Comparison of statistics
Statistics of probabilistic model

Let us try another model – the bigram model

- For every character

(Simulated) probabilistic model →

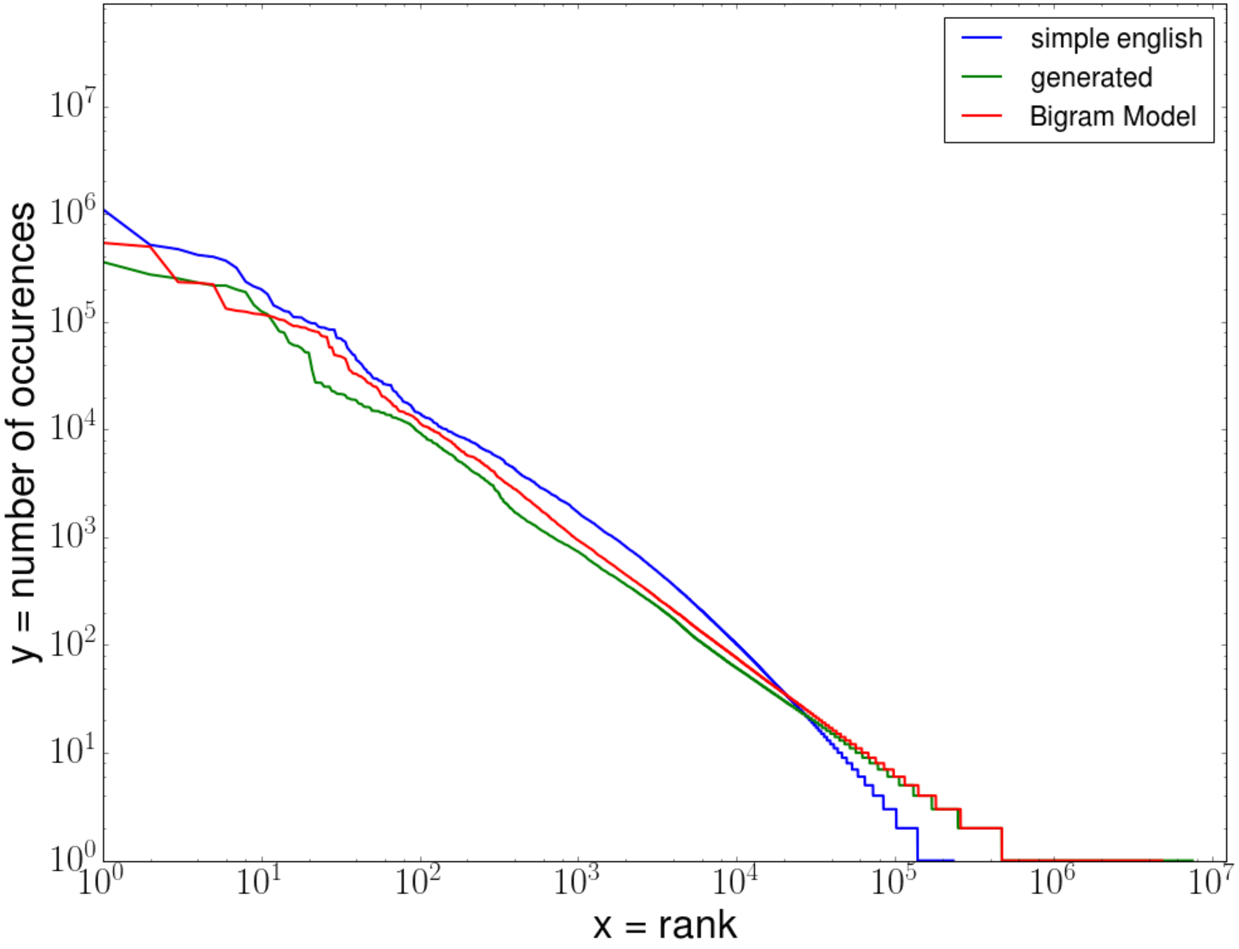
  – Learn a unigram distribution which contains the likelihood for the next character.

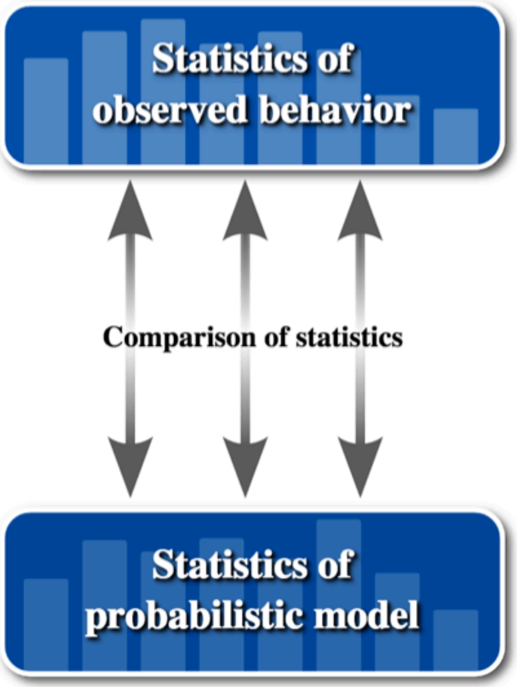
  – Draw the next character from this distribution

- This disregards the length distribution

Bigram model seems closer in the plot

Word frequencies depending on word rank on (Simple) Englisch Wikipedia



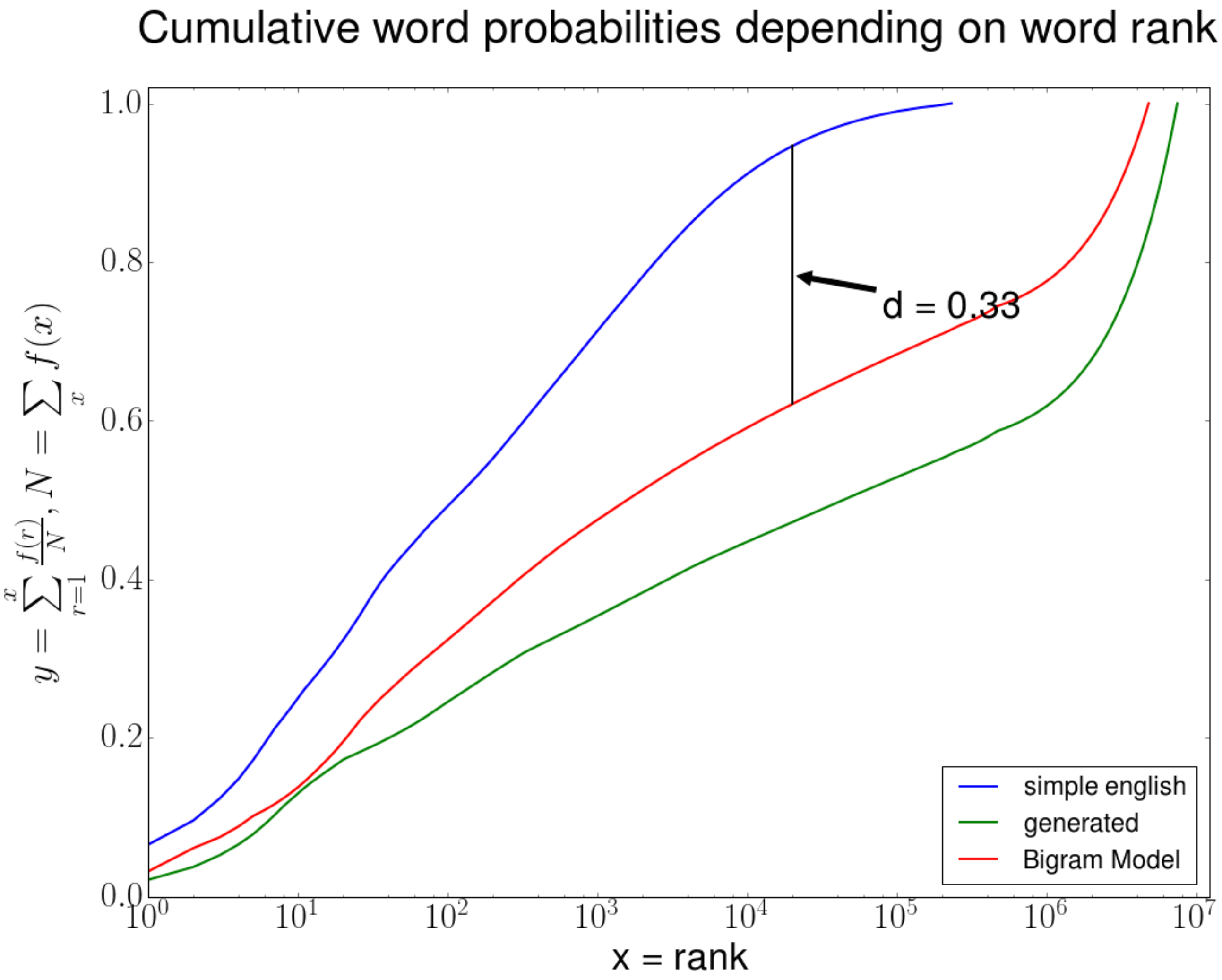


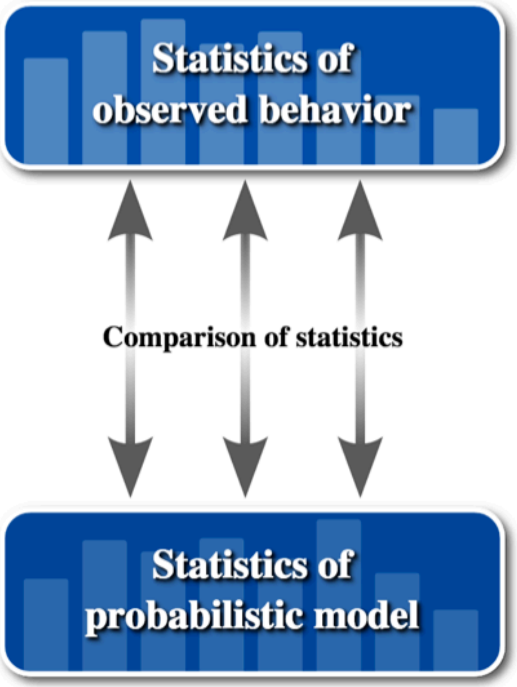
Statistics of observed behavior

Comparison of statistics

Statistics of probabilistic model

Cumulative plot says the same!

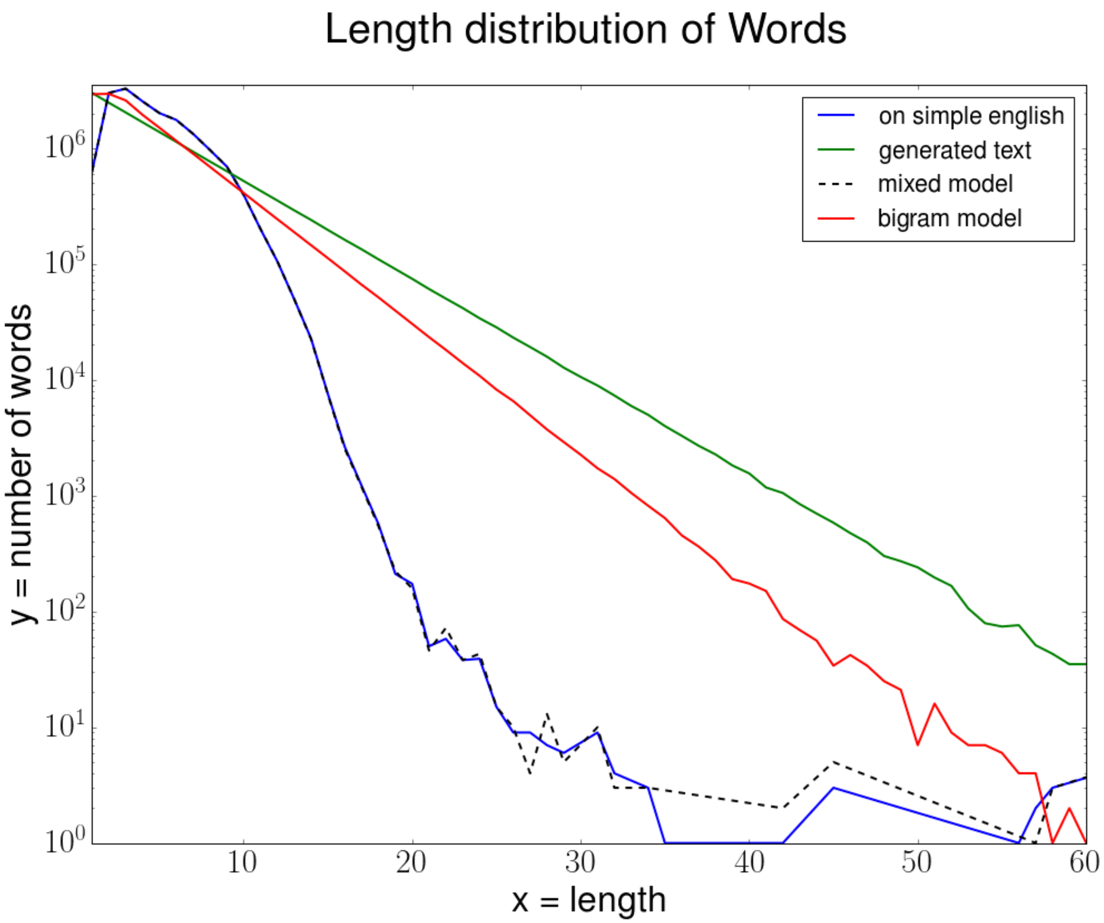



Statistics of observed behavior
Comparison of statistics
Statistics of probabilistic model

Comparing the length distribution

- bigram model still falls exponentially
  - Though for n=2 it fits exactly

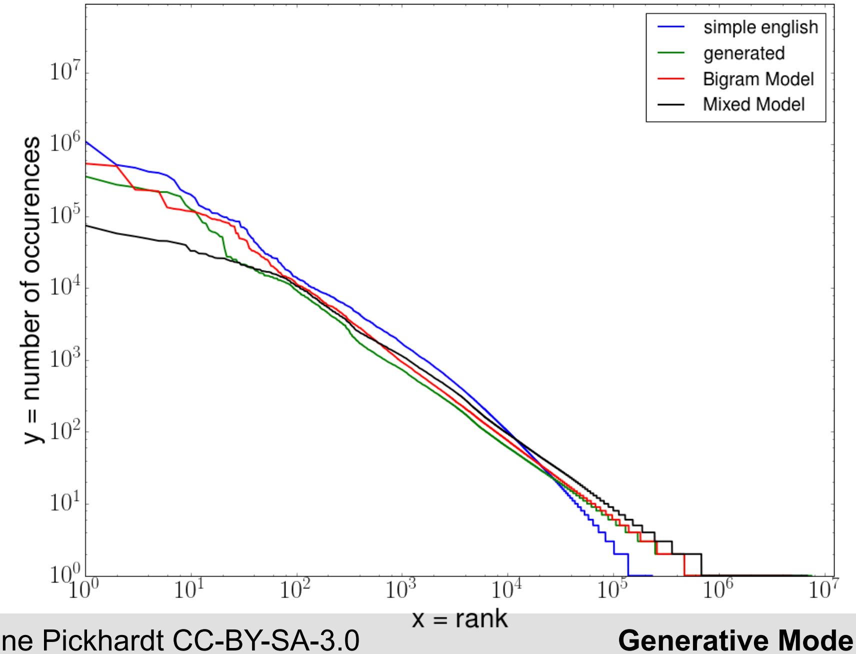
- Mixed model obviously follows original length distribution



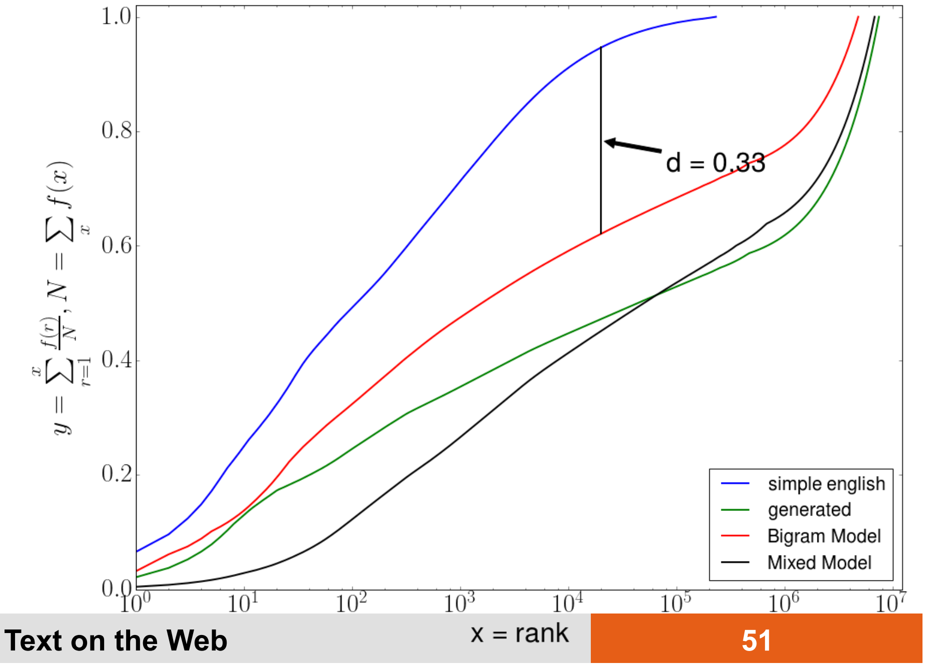
Comparing all 3 generative models

- All models are still far from being close to the observed data
- More sophisticated models tend to be closer.
- Goal is always to find small, close models

Word frequencies depending on word rank on (Simple) Englisch Wikipedia



Cumulative word probabilities depending on word rank



What can we explain now?

- Remember a reason to build generative models was to explain how or why something is in the way it is.
- We might say that the zipf distribution of words come from the character distribution (which was also Zipf)
- More model parameters yield better approximations
- Will they also explain more?
  - Not clear parameters have to be explained

Thank you for your attention!

Contact:
Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

Copyright: