Lesson5:
Generative Models for Text on the Web
Unit3:
Evaluating the Quality of a generative Model

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

WeST
People and Knowledge Networks

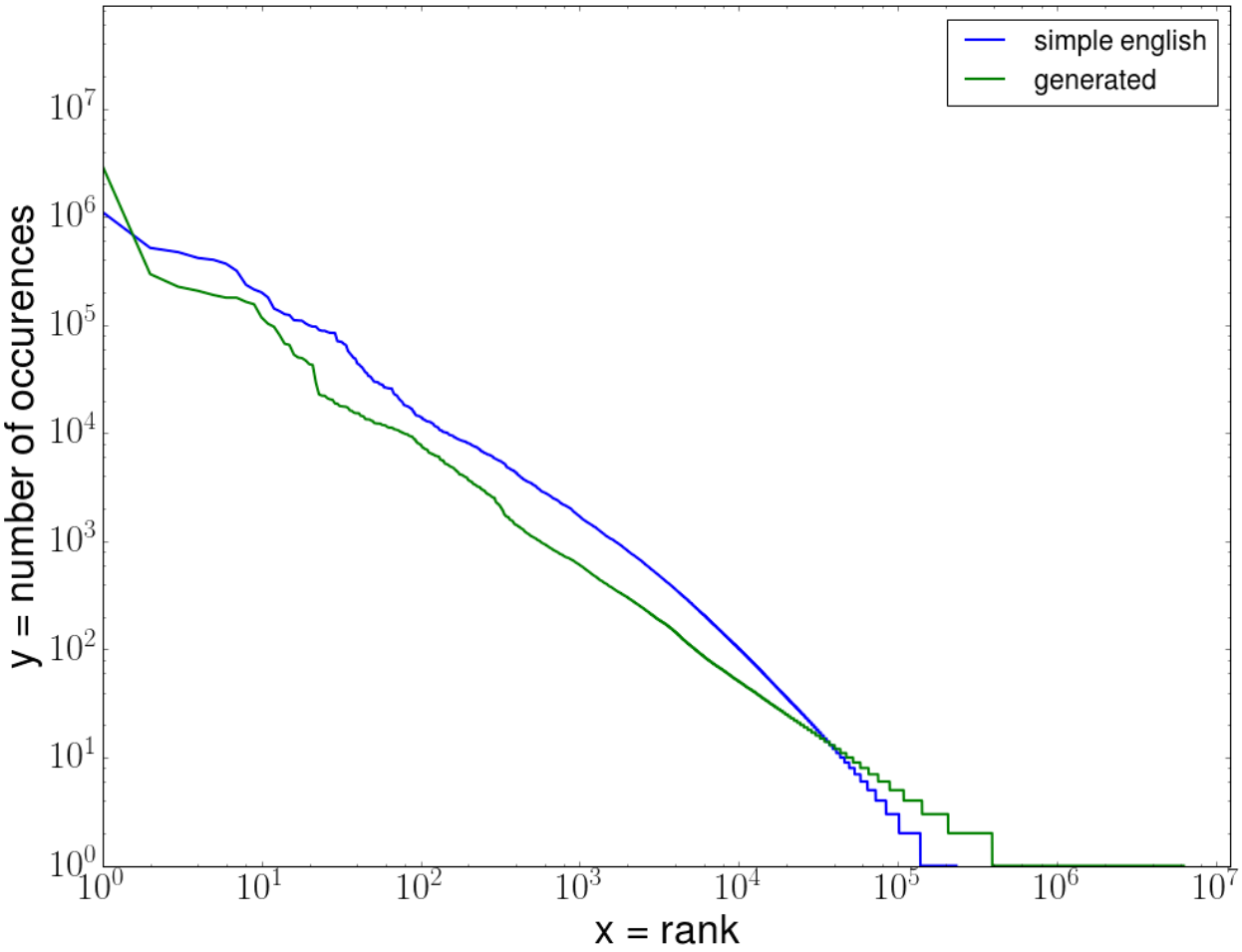
Institute for Web Science and Technologies · University of Koblenz-Landau, Germany

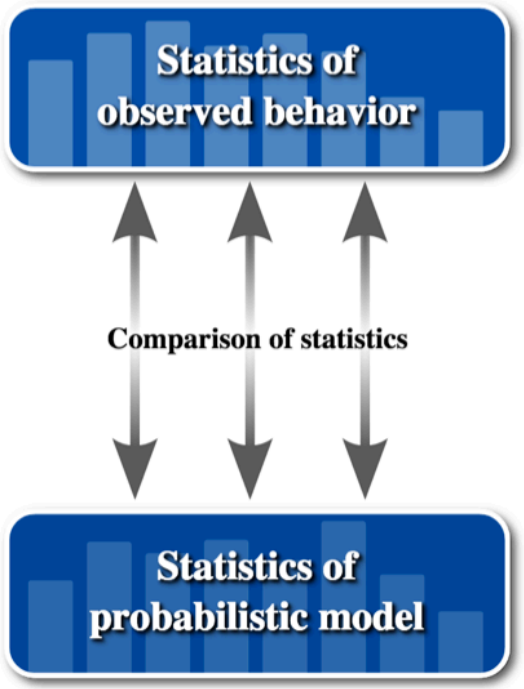
Completing this unit you should

- See that it makes sense to compare statistics
- Understand that comparing statistics is not a well defined task
- Be aware of the fact that very different models could lead to the same statistics

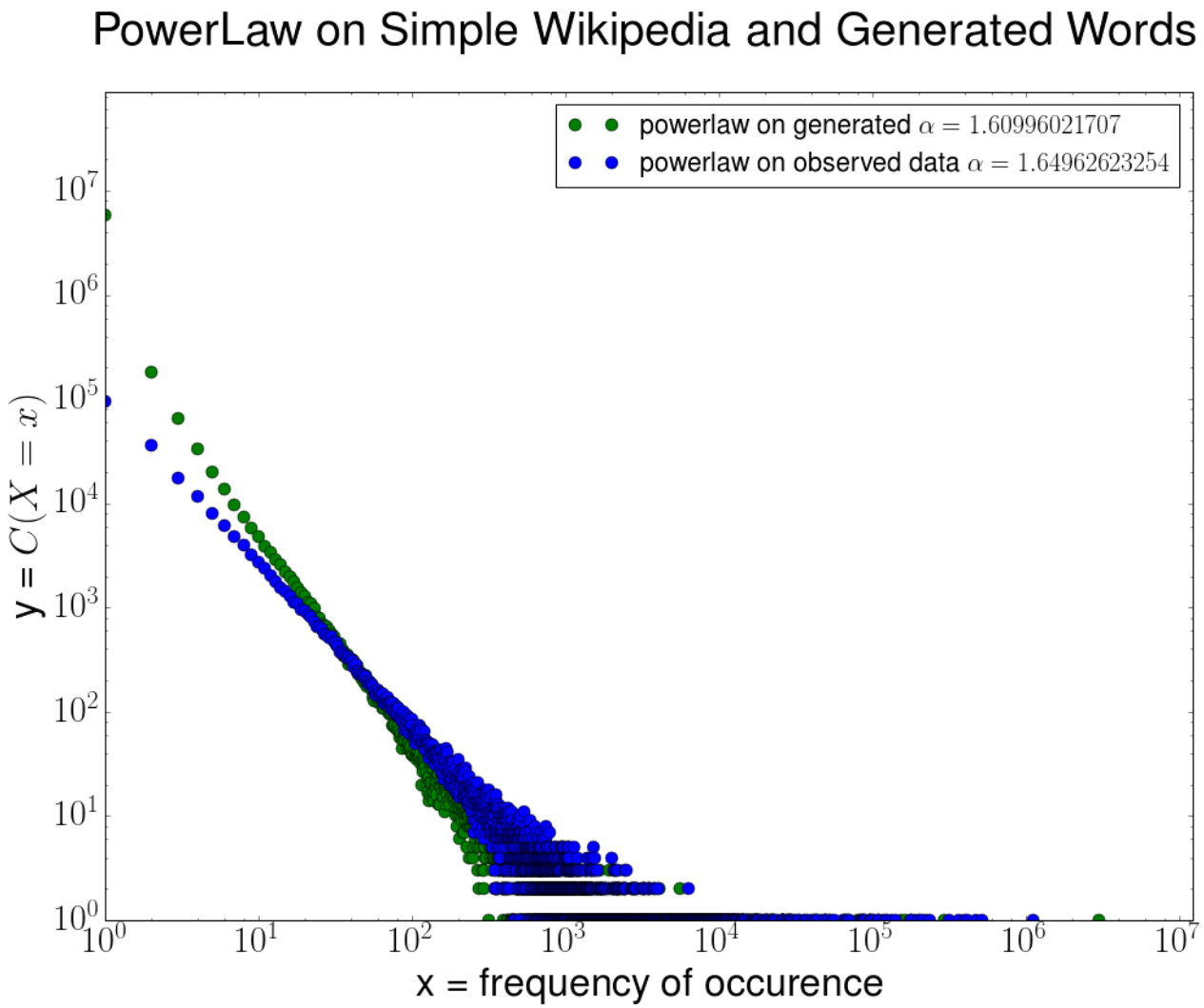
Plotting the Zipf distribution

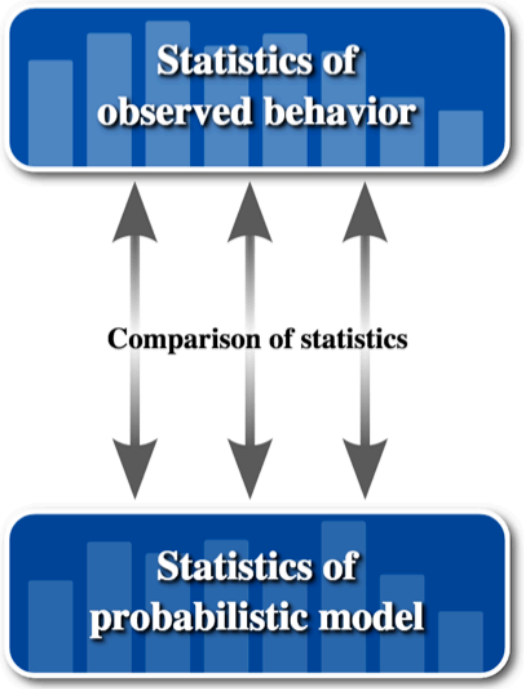
Word frequencies depending on word rank on (Simple) Englisch Wikipedia





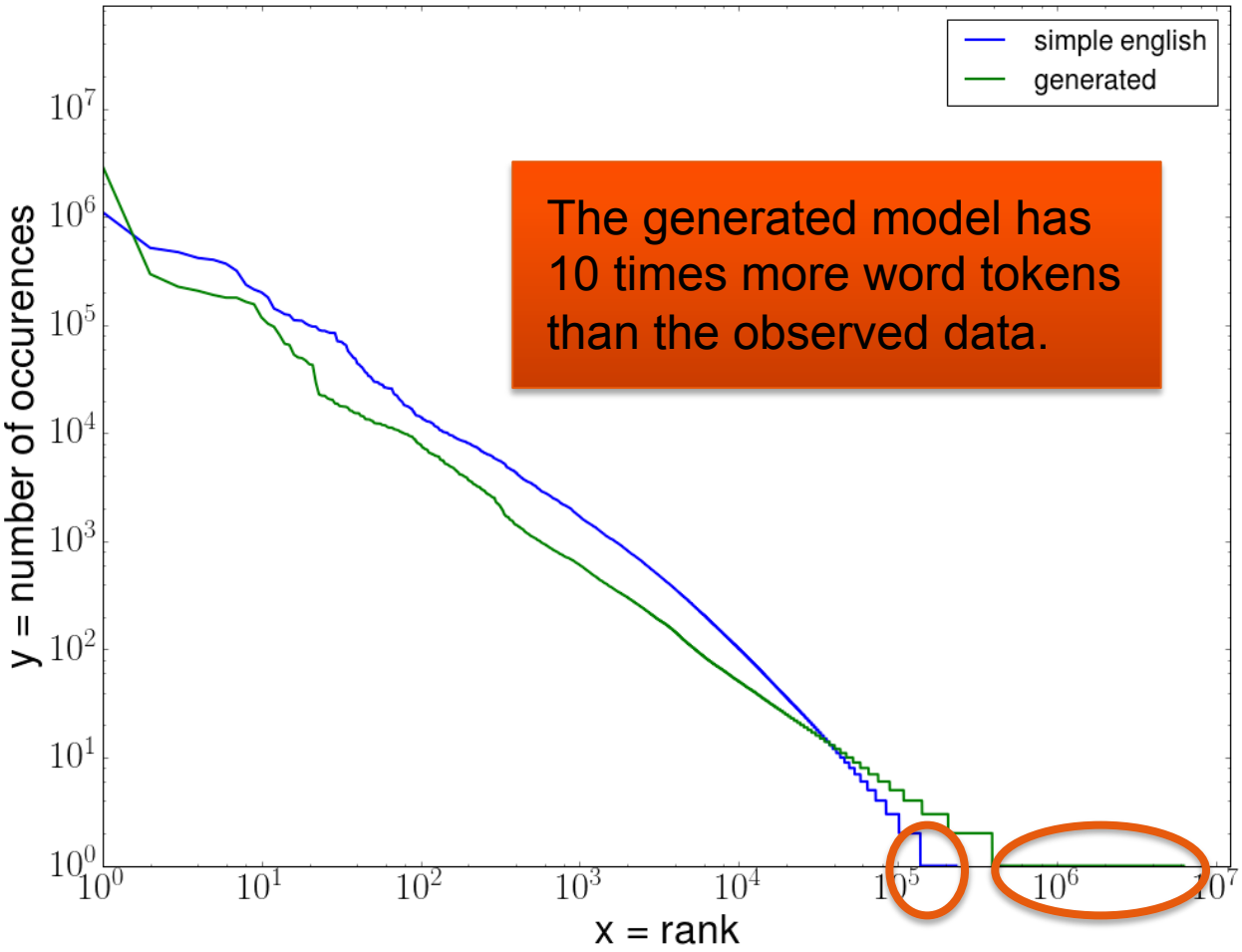
Powerlaw plot and exponent look good



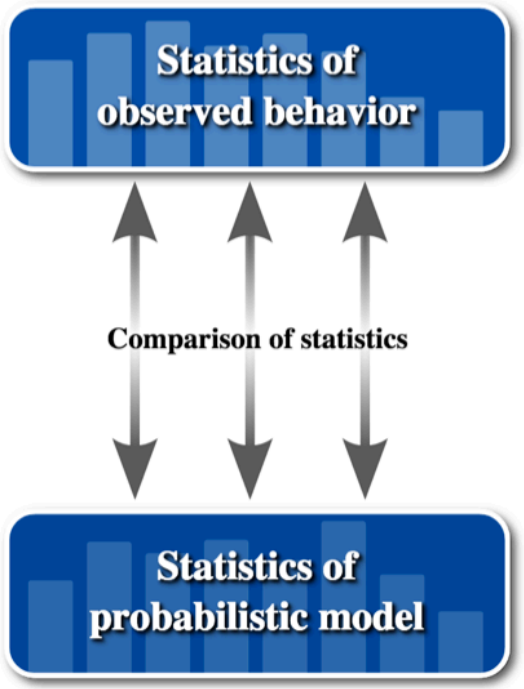


Remember not to get fooled by the log scale

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

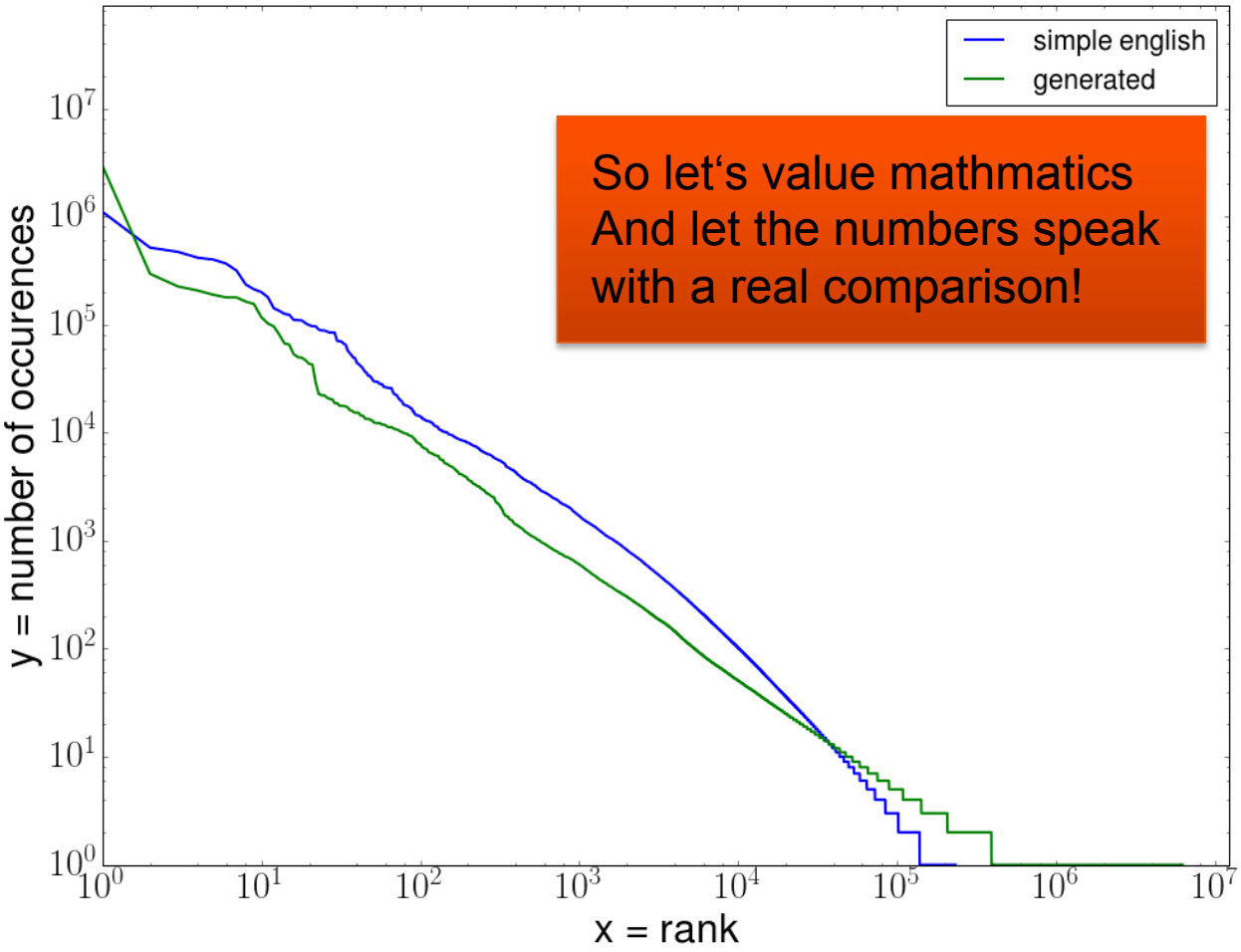


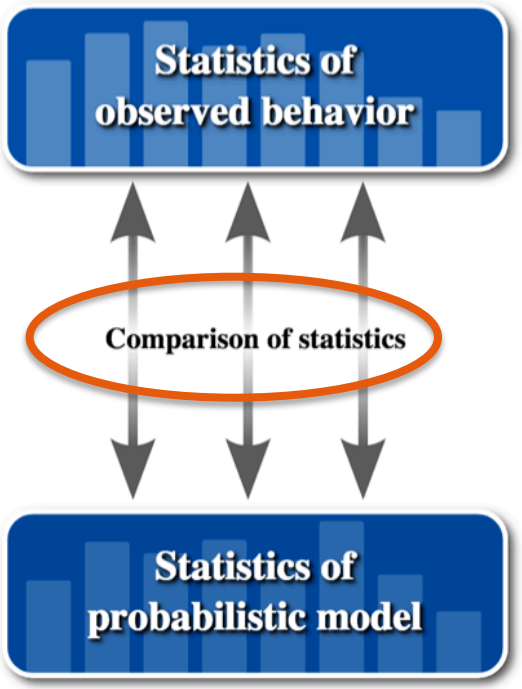
The generated model has 10 times more word tokens than the observed data.



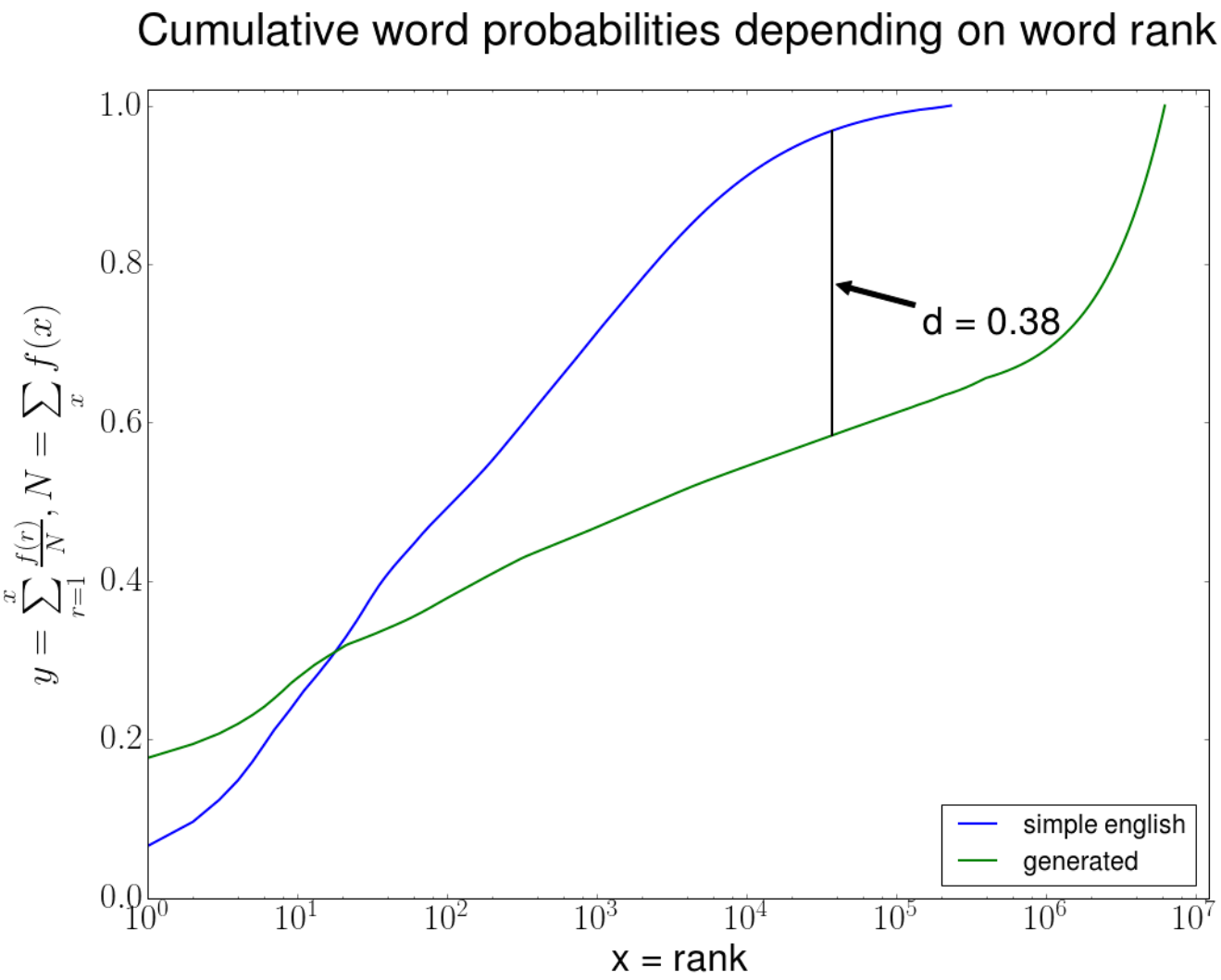
Who remembers a good distance measure?

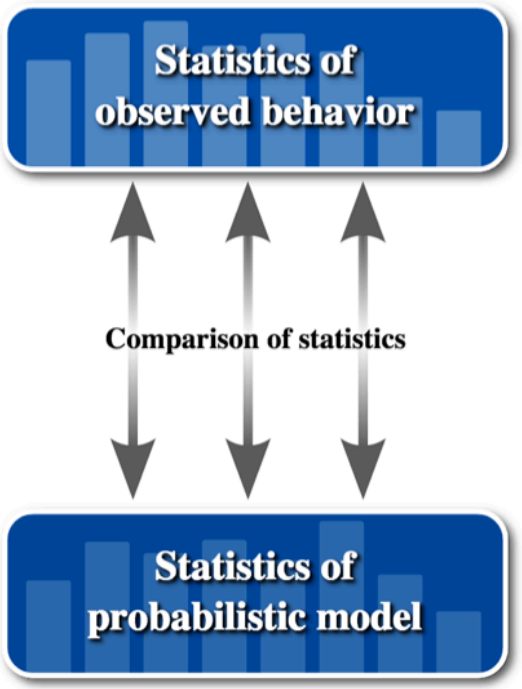
Word frequencies depending on word rank on (Simple) Englisch Wikipedia



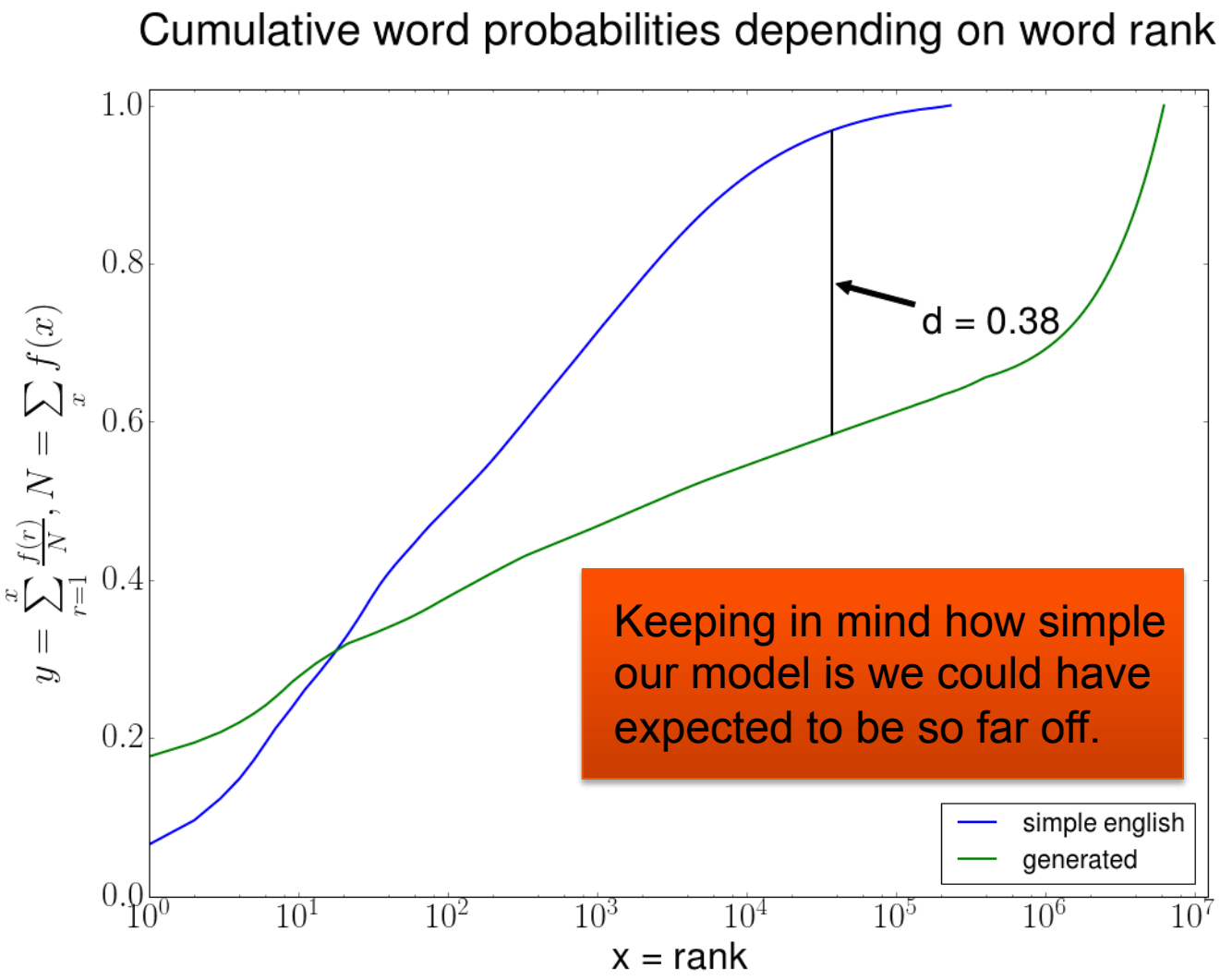


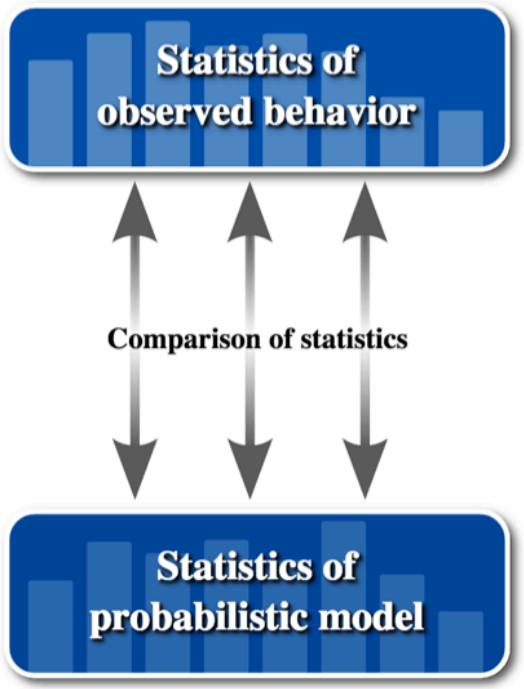
OUCH! That's a surprise – Or isn't it?



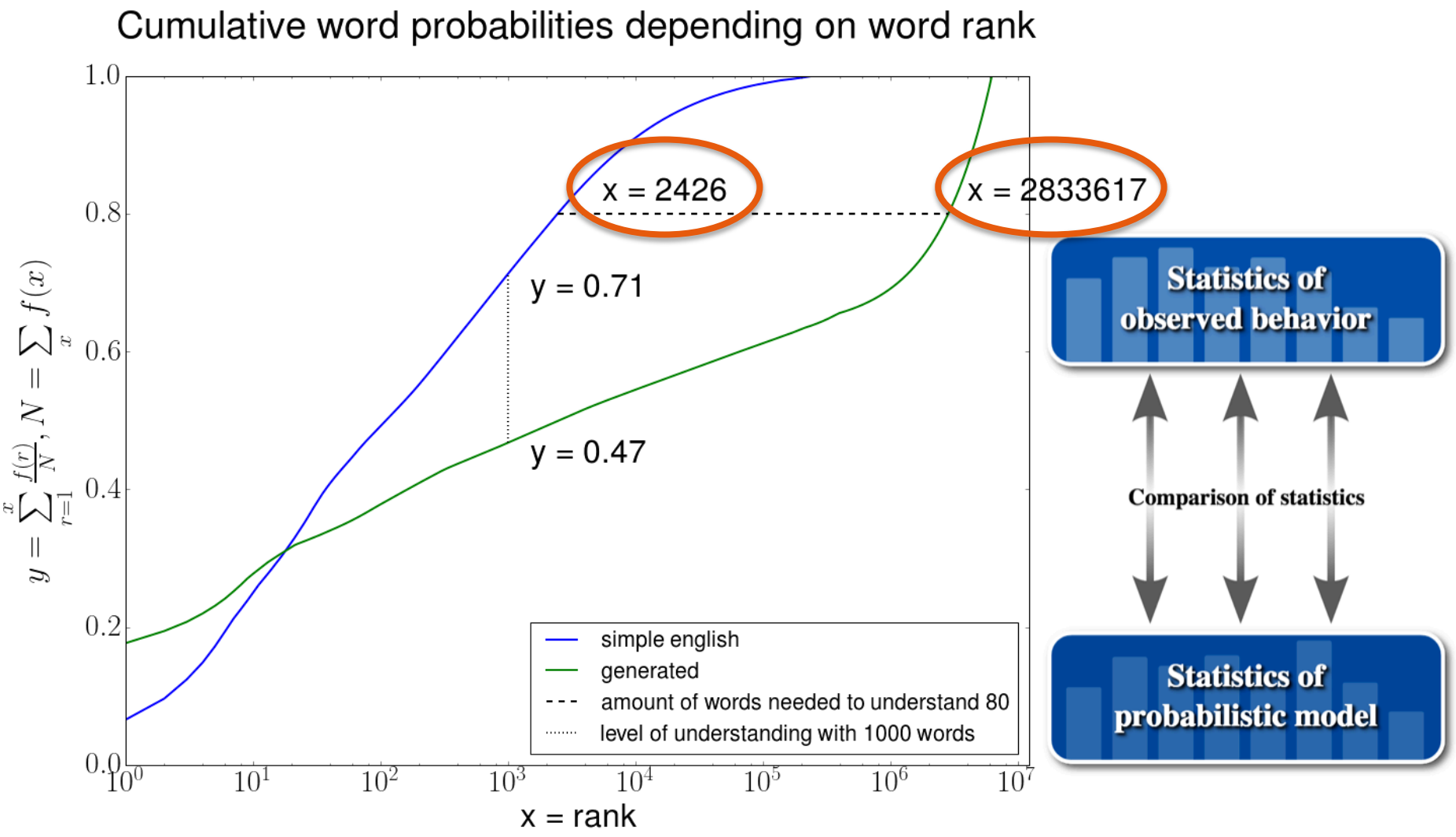


OUCH! That's a surprise – Or isn't it?

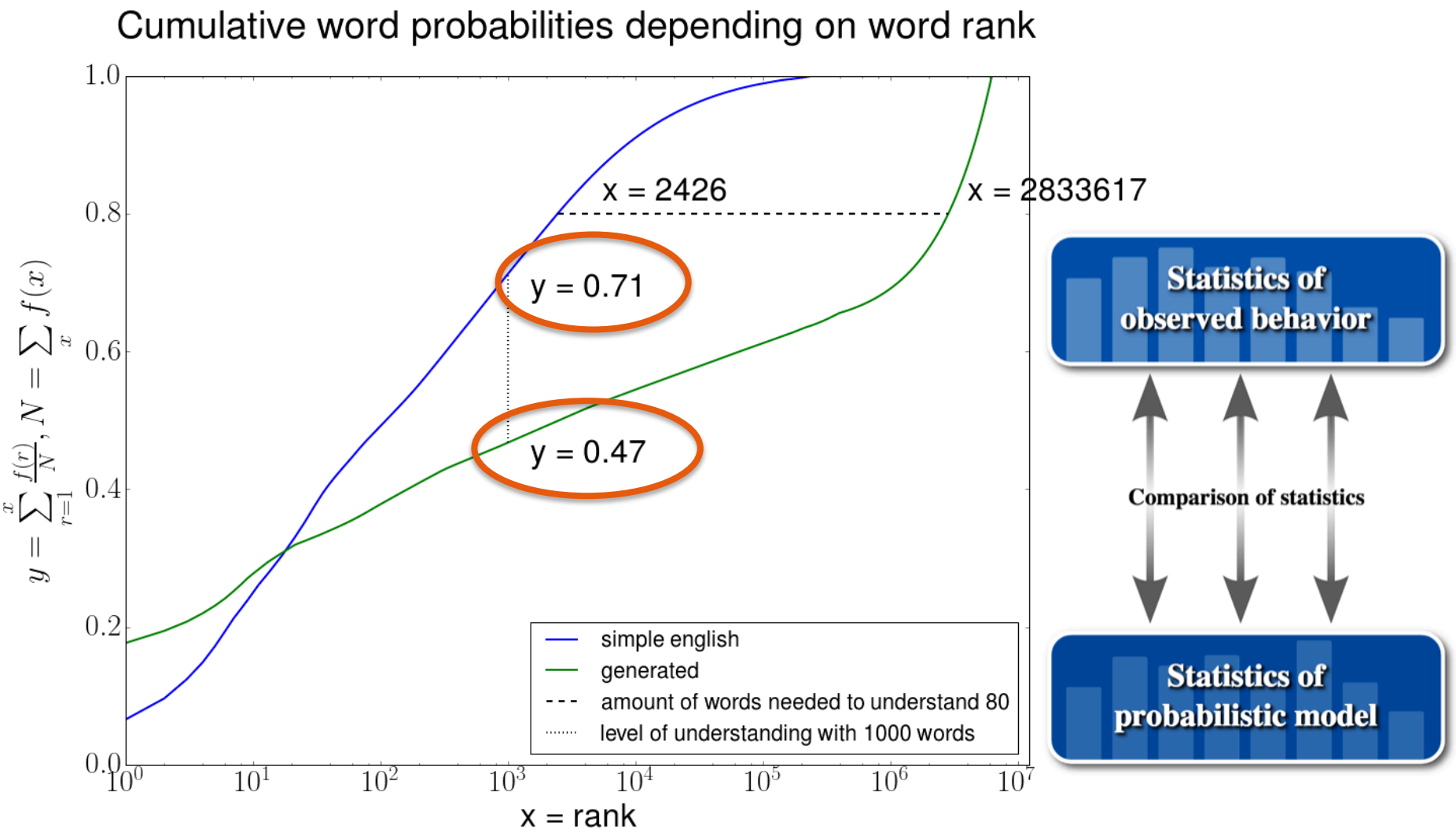




To understand 80% you need to know 1150 times more words in the generated language

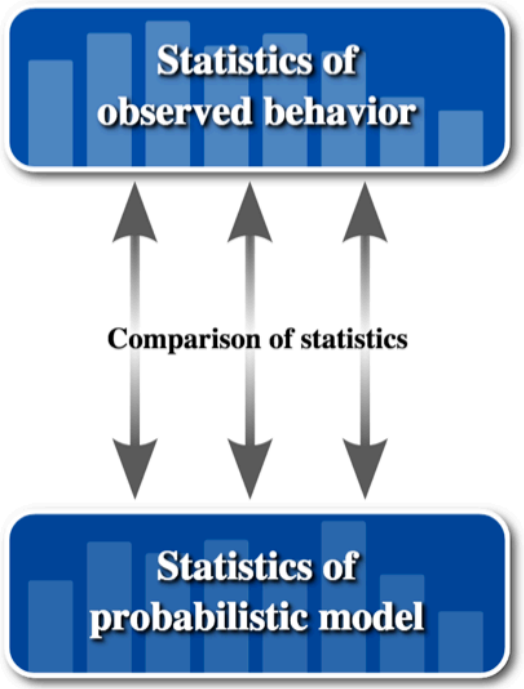


Level of expressiveness with 1000 known words also differs drastically

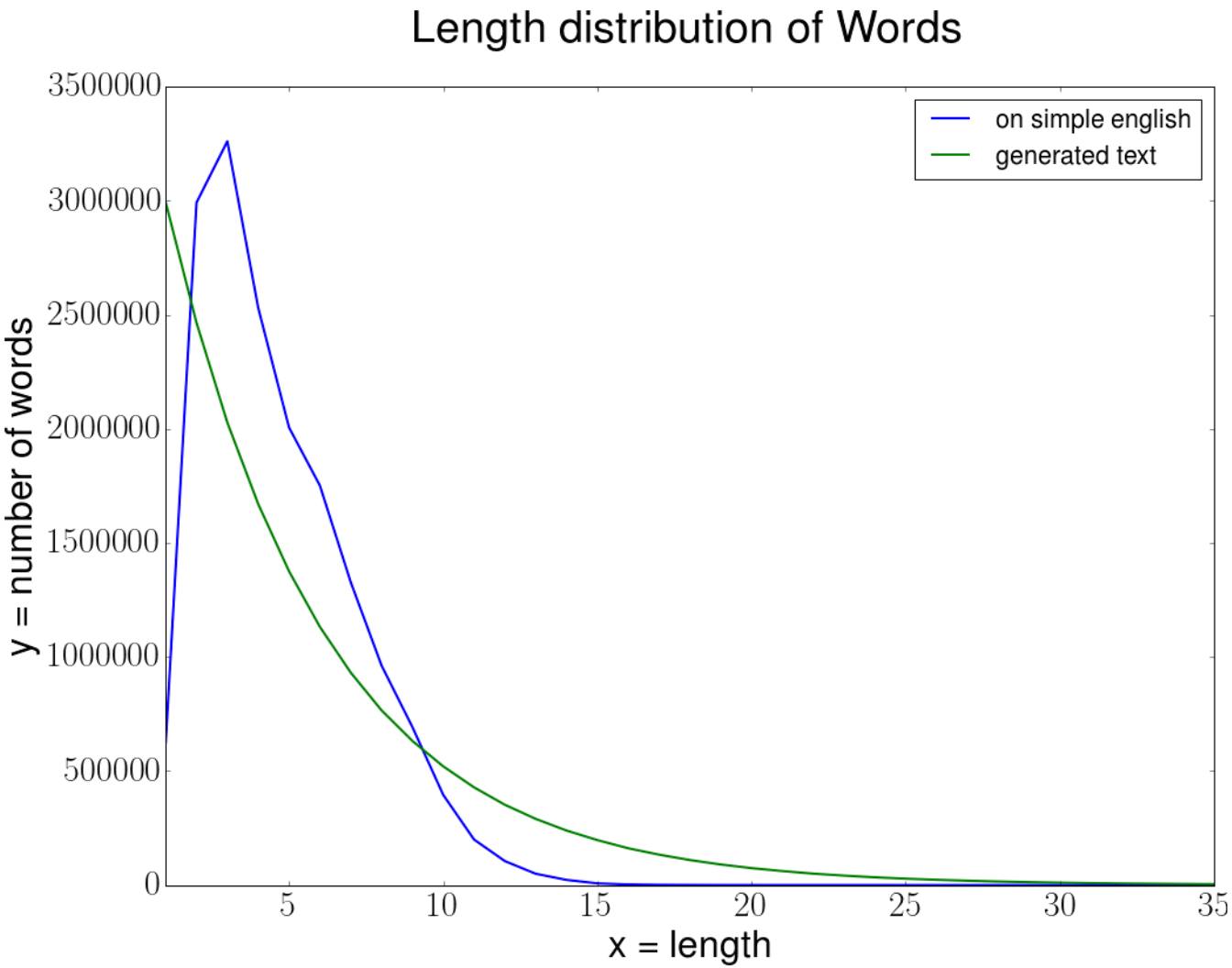


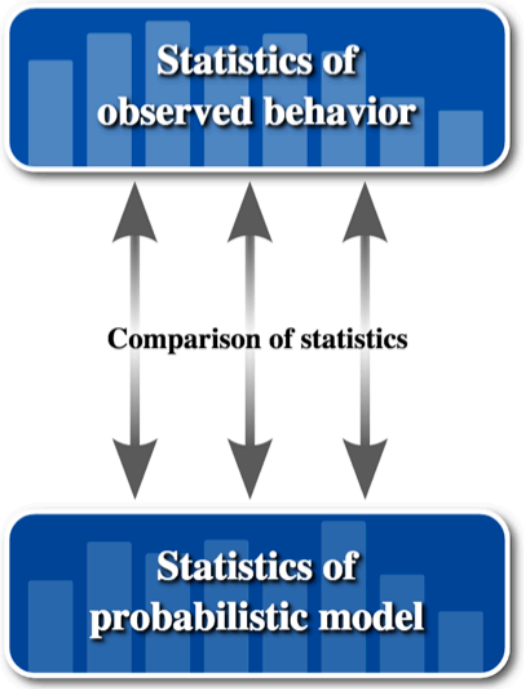
More descriptive statistics for text modelling

- Word length distribution
- Document frequency
- Sentence distribution
- Number of documents
  - Can't be used in since we didn't generate documents

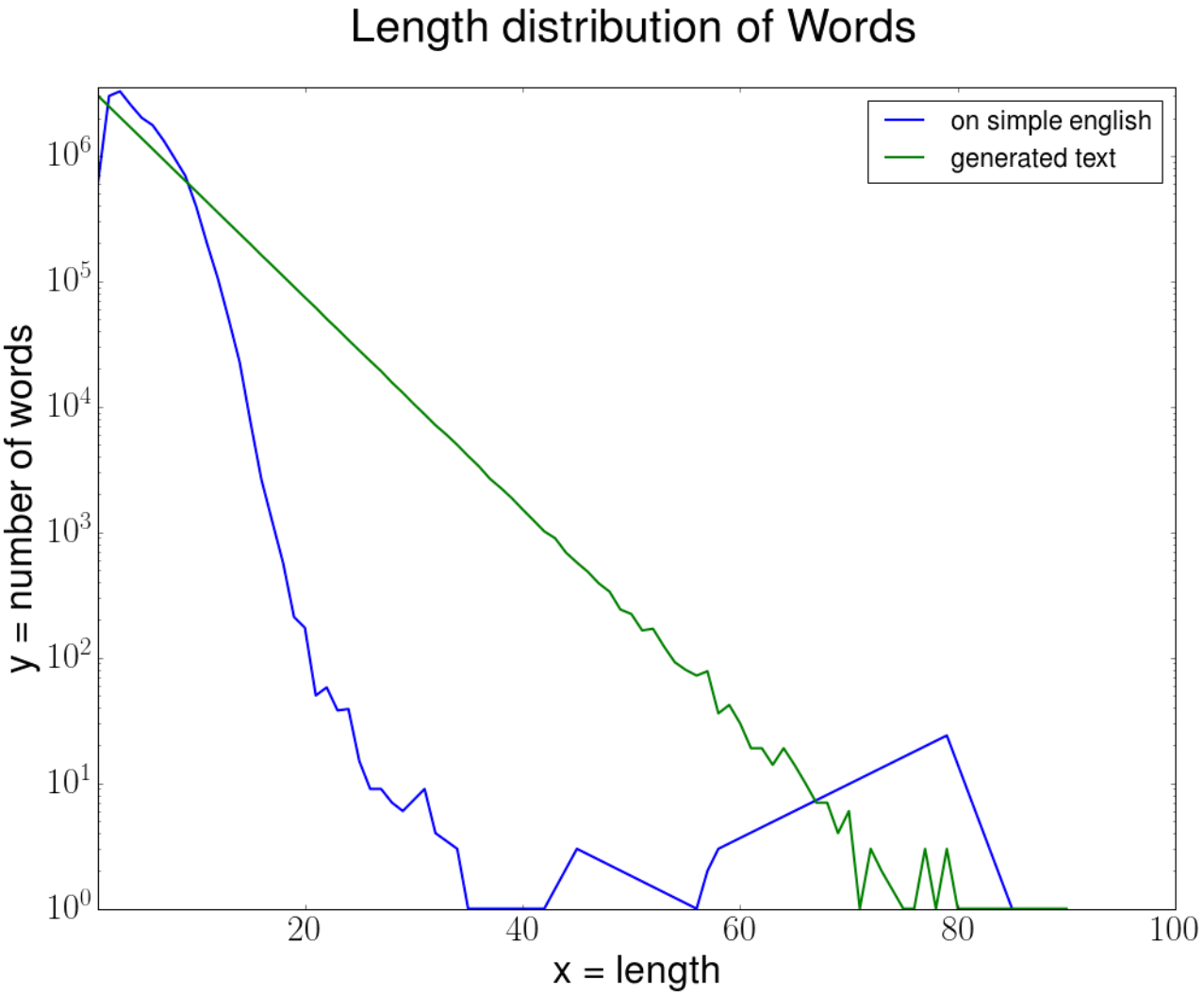


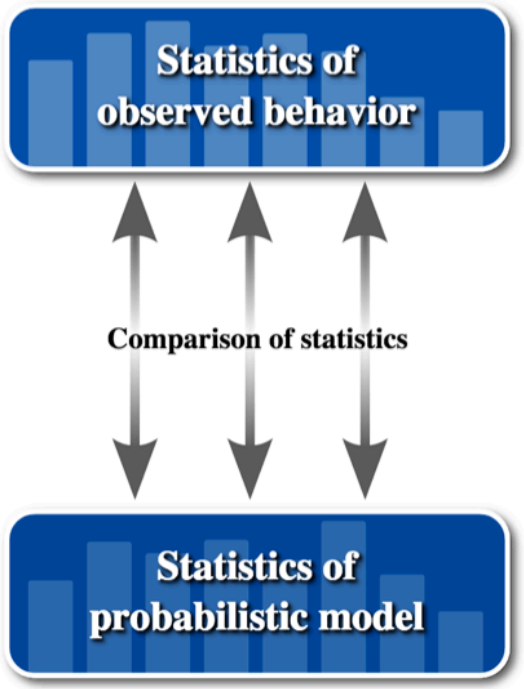
Generated text has too many short and long words. Too few between 3 and 9 characters



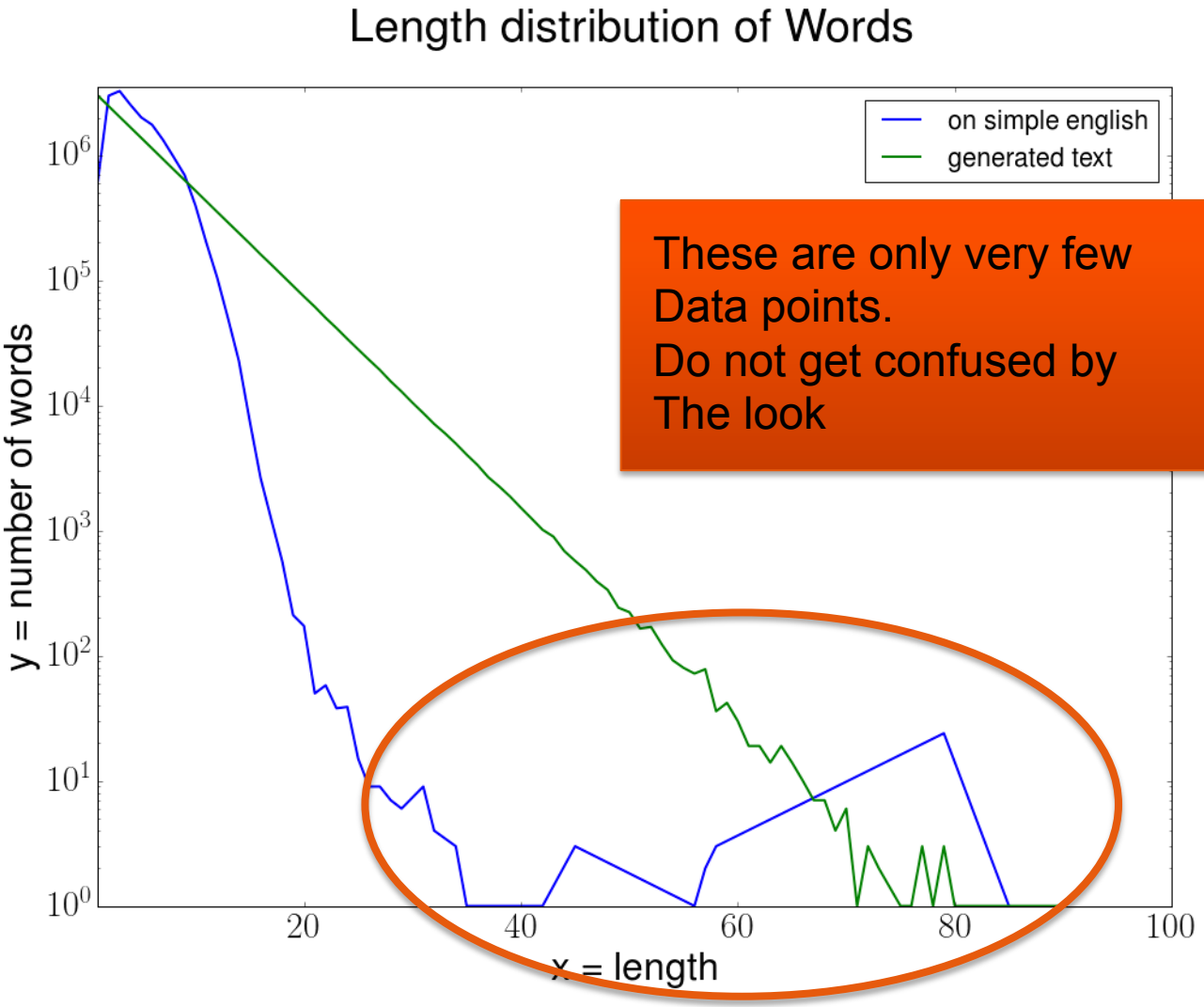


Generated curve is a perfect line on a plot with log y-axis. Exponential decrease!

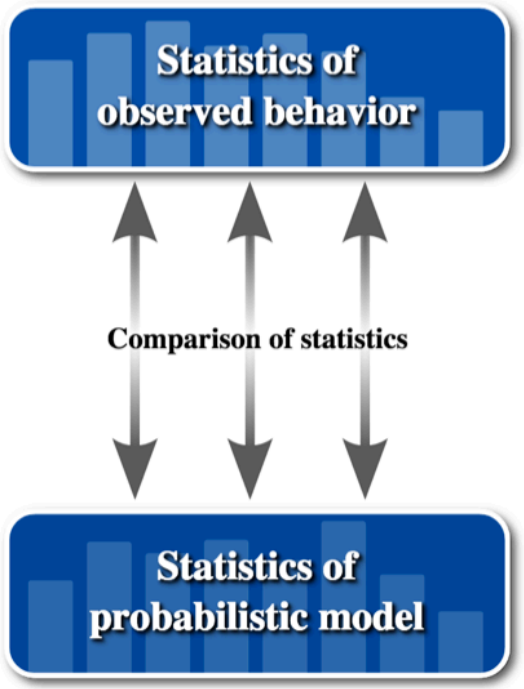




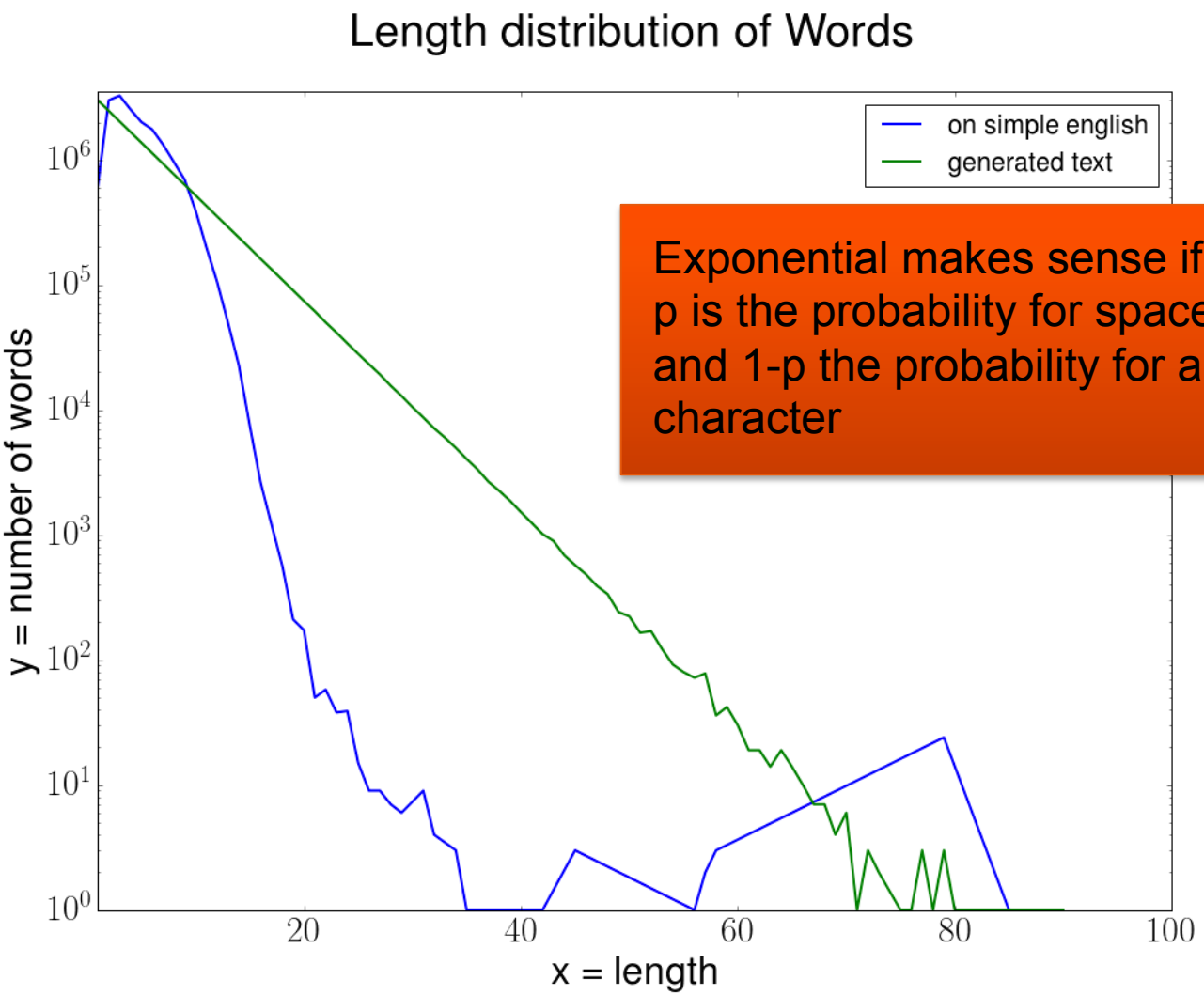
Generated curve is a perfect line on a plot with log y-axis. Exponential decrease!



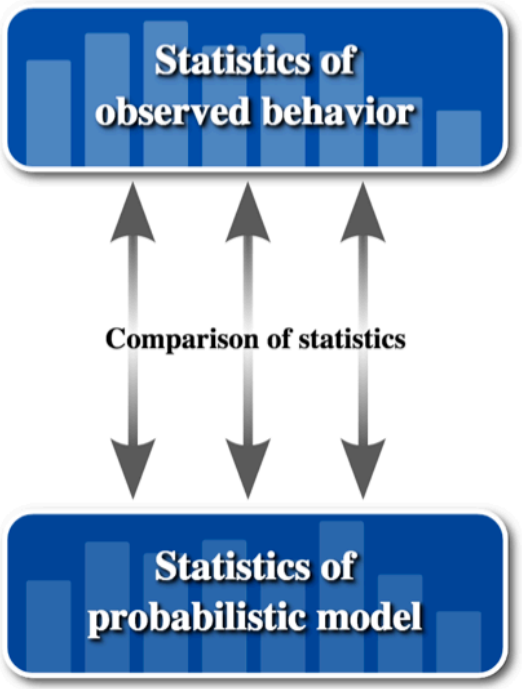
These are only very few Data points.
Do not get confused by The look



$$P(\texttt{len(word)} = n) = (1-p)^n p \sim c^n, 0 < c < 1$$



Exponential makes sense if p is the probability for space and 1-p the probability for a character



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: