Lecture 2:
Words and Basic Text Processing

CS 585, Fall 2016

Introduction to Natural Language Processing

http://people.cs.umass.edu/~brenocon/inlp2016

Brendan O'Connor

College of Information and Computer Sciences
University of Massachusetts Amherst

Announcements

- Currently: small assignments
  - HW0 due tomorrow
  - HW1 -- word counting programming (out soon)
  - HW2 -- n-gram language modeling (next week)

- Video link
- Project info on website (poster session!...)

- Collaboration policy (different than what I said briefly in class last time)
  - All of the content you submit, both code and text, needs to be produced independently.
  - You may discuss problems. List your collaborators you worked with.
  - Do NOT share code or written materials.
  - Cite sources.
- Course website has more complete version.

Today

- Python demo
- Basic text processing: Regular expressions
- Word counts

Python

- This weekend: make sure you can run Python
  - Recommended: Anaconda Python https://www.continuum.io/downloads
  - Python 2.7
  - IPython Notebook http://ipython.org/notebook.html
- Python interactive interpreter
- Python scripts

- Regular expressions (other slides)

Text normalization

- Every NLP task needs text normalization
  - 1. Segment/tokenize words in running text
  - 2. Normalizing word formats
  - 3. Sentence segmentation (typically)

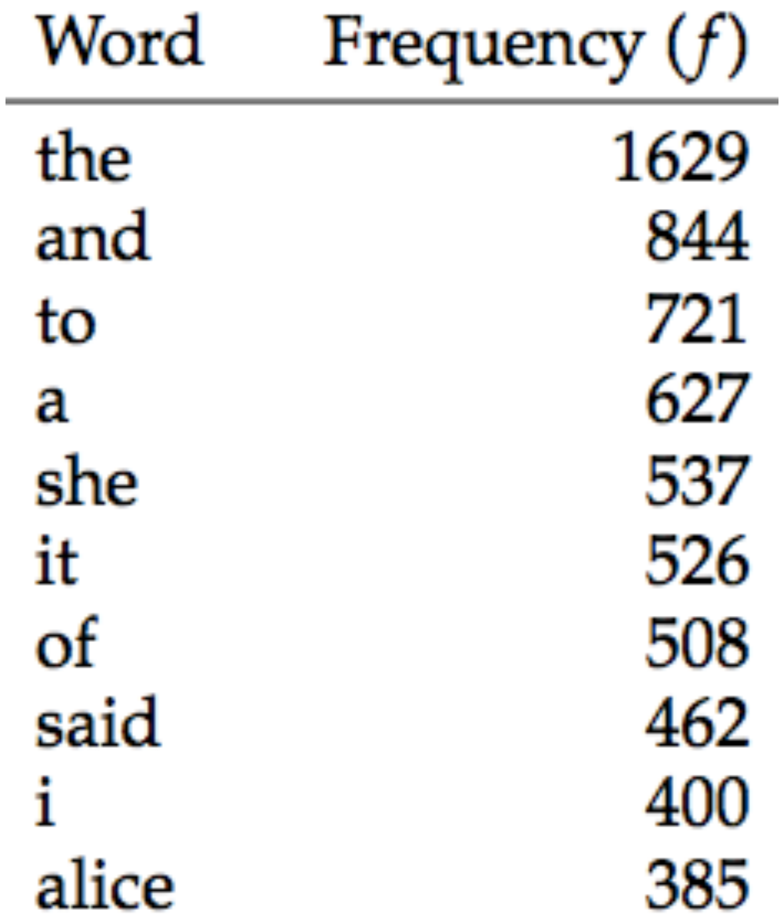
Type vs Token

- I saw one cat and then more cats!

- **N** = number of tokens
- **V** = vocabulary = set of types

	Tokens = N	Types = |V|
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

Word frequencies



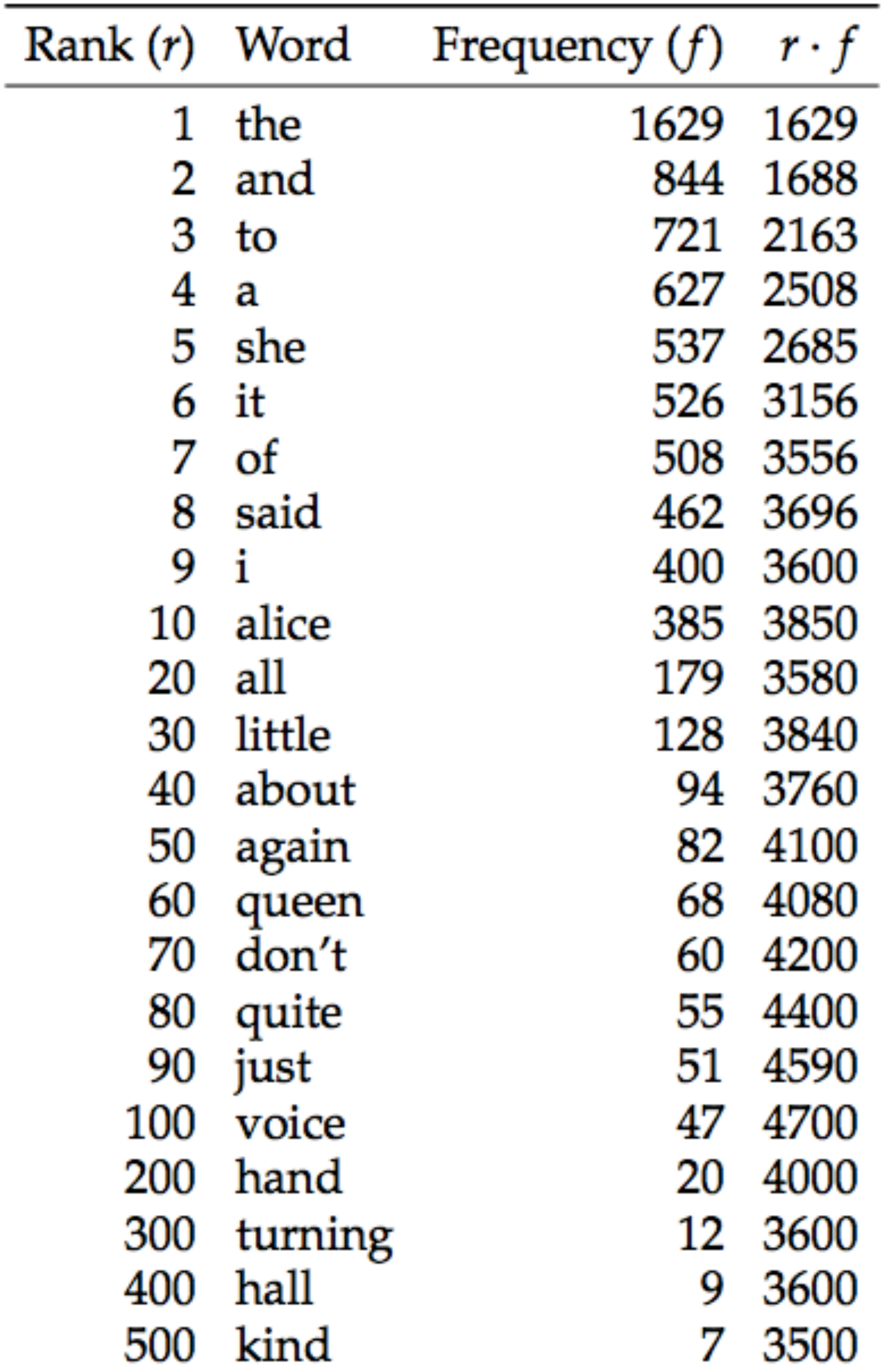
Word	Frequency (f)
the	1629
and	844
to	721
a	627
she	537
it	526
of	508
said	462
i	400
alice	385

Alice's Adventures in Wonderland, by Lewis Carroll

Zipf’s Law

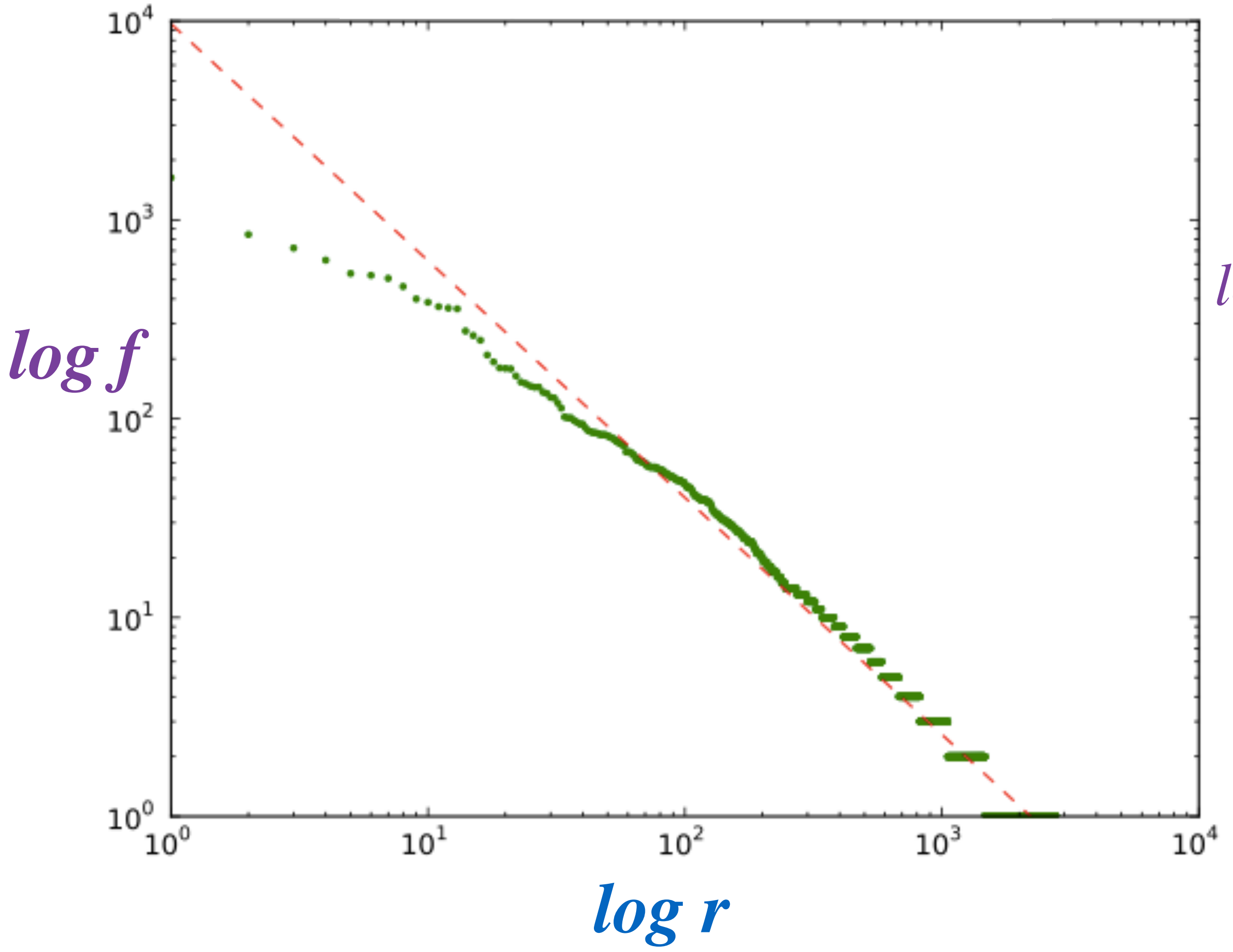
- When word types are ranked by frequency, then frequency (f) * rank (r) is roughly equal to some constant (k)

$$f \times r = k$$



Rank (r)	Word	Frequency (f)	$r \cdot f$
1	the	1629	1629
2	and	844	1688
3	to	721	2163
4	a	627	2508
5	she	537	2685
6	it	526	3156
7	of	508	3556
8	said	462	3696
9	i	400	3600
10	alice	385	3850
20	all	179	3580
30	little	128	3840
40	about	94	3760
50	again	82	4100
60	queen	68	4080
70	don’t	60	4200
80	quite	55	4400
90	just	51	4590
100	voice	47	4700
200	hand	20	4000
300	turning	12	3600
400	hall	9	3600
500	kind	7	3500

Plot: log frequencies



recall:

$$f * r = k$$

$$\log f + \log r = \log k$$