# ANLP Week 1 / Lecture 2
# Thinking about Ambiguity and Words

Adam Lopez

(based on slides by Sharon Goldwater)

School of informatics

# This Lecture

Ambiguity

- What are different sources of ambiguity?
- When and how is ambiguity resolved?

Words and their distribution

- What are word types and tokens, and what is the characteristic frequency distribution of word tokens?
- What aspects of frequency distributions of words are similar between languages, and what aspects are different, and why?

# Thinking More About Ambiguity

# Reminder: What Makes Language Difficult?

In Lecture 1, I mentioned three characteristics:

- Ambiguities on many levels, need context to disambiguate
- Rules, but many exceptions
- Language is infinite. We cannot see examples of everything, and the vast majority of what we do see is rare

# Revisiting Ambiguity

I discussed two jokes that use different types of ambiguity:

1. Lexical semantic ambiguity (meaning of a word):

   *I'm not a fan of the new pound coin, but then again, I hate all change.*[1]

2. Syntactic ambiguity (relationship between words):

   *One morning I shot an elephant in my pajamas. How he got in my pajamas I don't know.*[2]

[1] Ken Cheng, 2017. (Winner of Dave's Funniest Joke of the Fringe award.)
[2] Groucho Marx, in the 1930 film Animal Crackers.

# Ambiguity and Context

- Ambiguity is usually resolved by context or world knowledge.
  - Word-level ambiguity resolved by sentence-level context:
    The change of scenery was nice.
  - Phrase-level ambiguity resolved by sentence-level context:
    I ate the carrots in the garden, after I brought them inside.
  - Sentence-level ambiguity resolved by world knowledge:
    I cooked the fish in the freezer.
- Challenges for NLP are to resolve ambiguity both
  - Correctly: requires good models of language; and
  - Efficiently: requires good algorithms for processing.

# A quick poll

In a moment I'll show a question. Don't shout out the answer, just think to yourself. Then I'll ask for a show of hands.

# Poll: type of ambiguity

What type of ambiguity does the following sentence contain?

I passed the bar this morning.

1. Lexical
2. Syntactic
3. Both
4. Neither
5. I don't know

# Activity: More Examples of Ambiguity

In a moment, I'll ask you to discuss some other examples.

1. I'll show some additional examples of ambiguity and ask you to think about them by yourself (2 minutes).
2. I'll ask you to talk to your neighbours in groups of 2-3 to see if you agree, or to help each other if you're stuck (4 minutes).
   - If you're all stuck, move on and ask afterwards.
3. When done, I'll ask a few volunteers to report back, and I can answer questions.

# Think to Yourself (2 minutes)

Look over the sentences below. For each one, try to:

- Identify **one** source of ambiguity or potential ambiguity. Is it lexical? Syntactic? Neither?
  - Some sentences may have multiple ambiguities! Just pick one!
- Decide if the ambiguity is already resolved by world knowledge or context. If so, which?

1. I like the other chair better.
2. I drew the girl with the jumper.
3. The first line of this joke:
   **Sam:** We should replace the sofa.
   **Alex:** Really? I wouldn't like being sat on all the time.

# Small Group Task (4 minutes)

- First, introduce yourselves if you haven't already.
- Then, each person pick one of the sentences below, and say:
  - What type of ambiguity did you find? If it is resolved by world knowledge or context, explain how.
  - If the sentence is still ambiguous, can you provide an unambiguous paraphrase (re-wording) of each plausible meaning?
- Do you all agree on the examples? Did you find anything that someone else missed?

1. I like the other chair better.
2. I drew the girl with the jumper.
3. The first line of this joke:
   **Sam:** We should replace the sofa.
   **Alex:** Really? I wouldn't like being sat on all the time.

# Recap: What Did You Find?

1. I like the other chair better.
2. I drew the girl with the jumper.
3. The first line of this joke:
   **Sam:** We should replace the sofa.
   **Alex:** Really? I wouldn't like being sat on all the time.

# Words as data
# (Types, tokens, and Zipf's law)

# Data: Words

In this class, we will consider **written language** (text). Keep in mind that writing is itself a technology!

What is a word? Possible definition: strings of letters separated by spaces

- But how about:
  - punctuation: commas, periods, etc are normally not part of words, but others less clear: high-risk, Joe's, @sloppyjoe
  - compounds: website, Computerlinguistikvorlesung
- And what if there are no spaces:
  伦敦每日快报指出,两台记载黛安娜王妃一九九七年巴黎死亡车祸调查资料的手提电脑,被从前大都会警察总长的办公室里偷走.

Processing text to decide/extract words is called **tokenization**.

# Word Counts

Out of 24m total word **tokens** (instances) in the English Europarl corpus, the most frequent are:

**any word**

| Frequency | Token |
|---|---|
| 1,698,599 | the |
| 849,256 | of |
| 793,731 | to |
| 640,257 | and |
| 508,560 | in |
| 407,638 | that |
| 400,467 | is |
| 394,778 | a |
| 263,040 | I |

**nouns**

| Frequency | Token |
|---|---|
| 124,598 | European |
| 104,325 | Mr |
| 92,195 | Commission |
| 66,781 | President |
| 62,867 | Parliament |
| 57,804 | Union |
| 53,683 | report |
| 53,547 | Council |
| 45,842 | States |

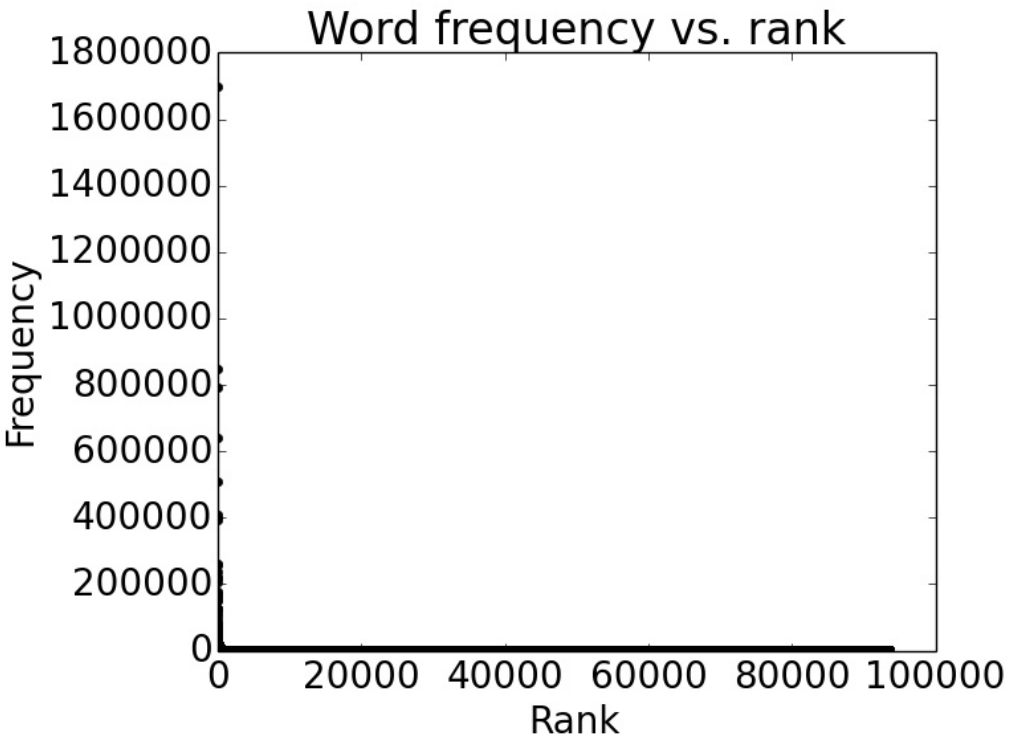
# Word Counts

But there are 93638 distinct words (**types**) altogether, and 36231 occur only once! Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a

# Plotting word frequencies

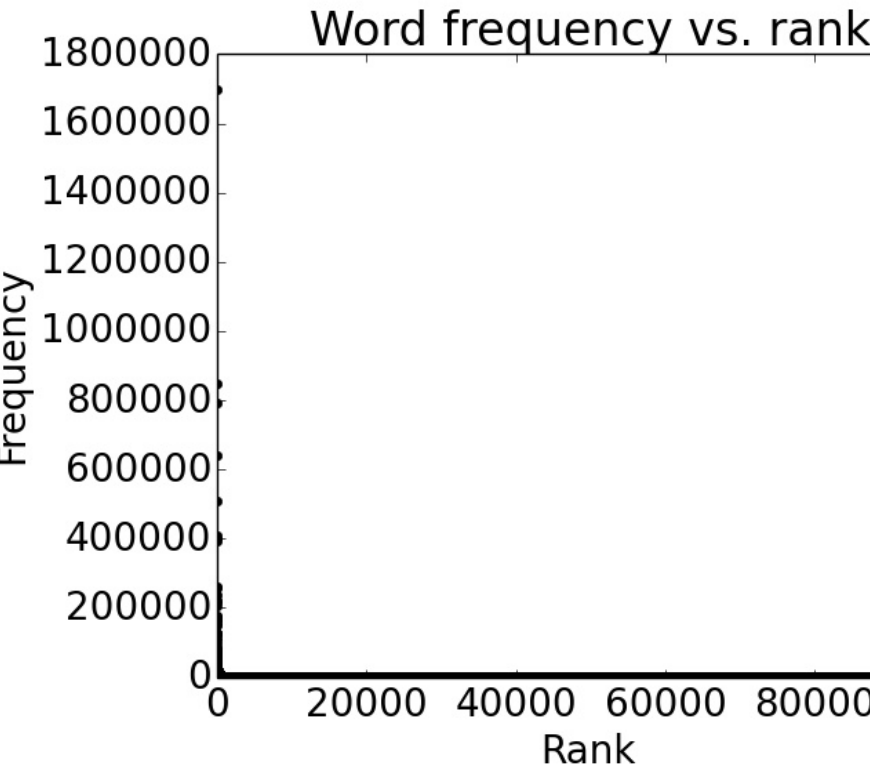
Order words by frequency. What is the freq of *n*th ranked word?

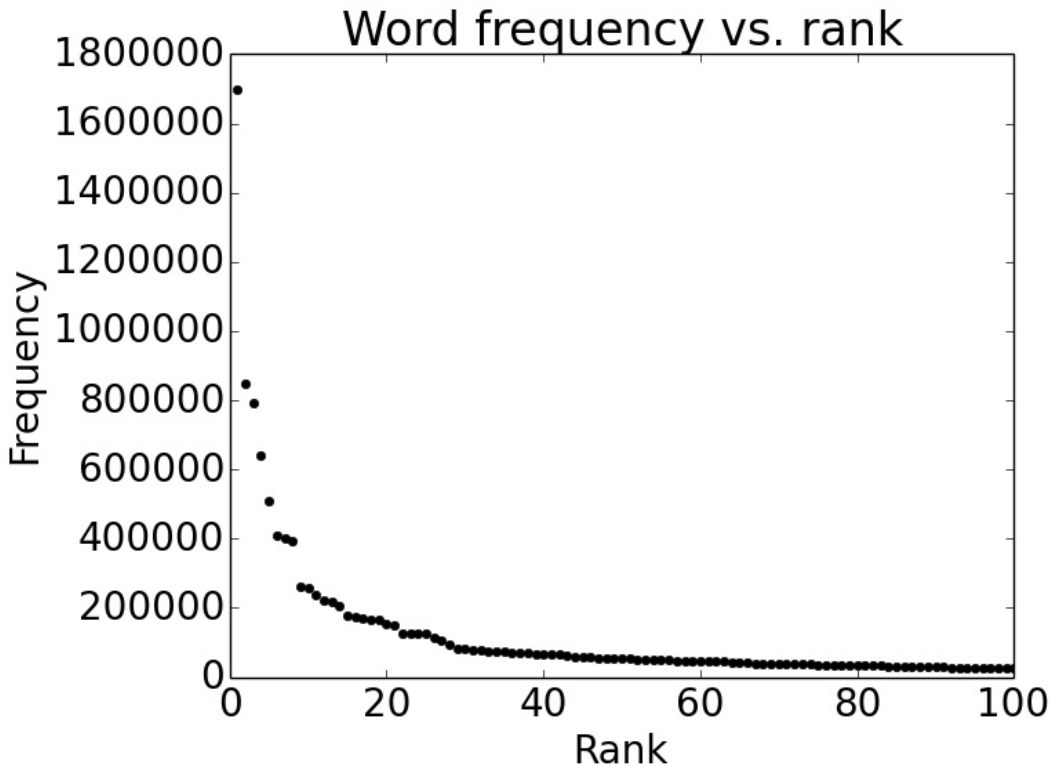



| Frequency | Token | Rank |
|---|---|---|
| 1,698,599 | the | 1 |
| 849,256 | of | 2 |
| 793,731 | to | 3 |
| 640,257 | and | 4 |
| 508,560 | in | 5 |
| 407,638 | that | 6 |
| 400,467 | is | 7 |
| 394,778 | a | 8 |
| 263,040 | I | 9 |

# Plotting word frequencies

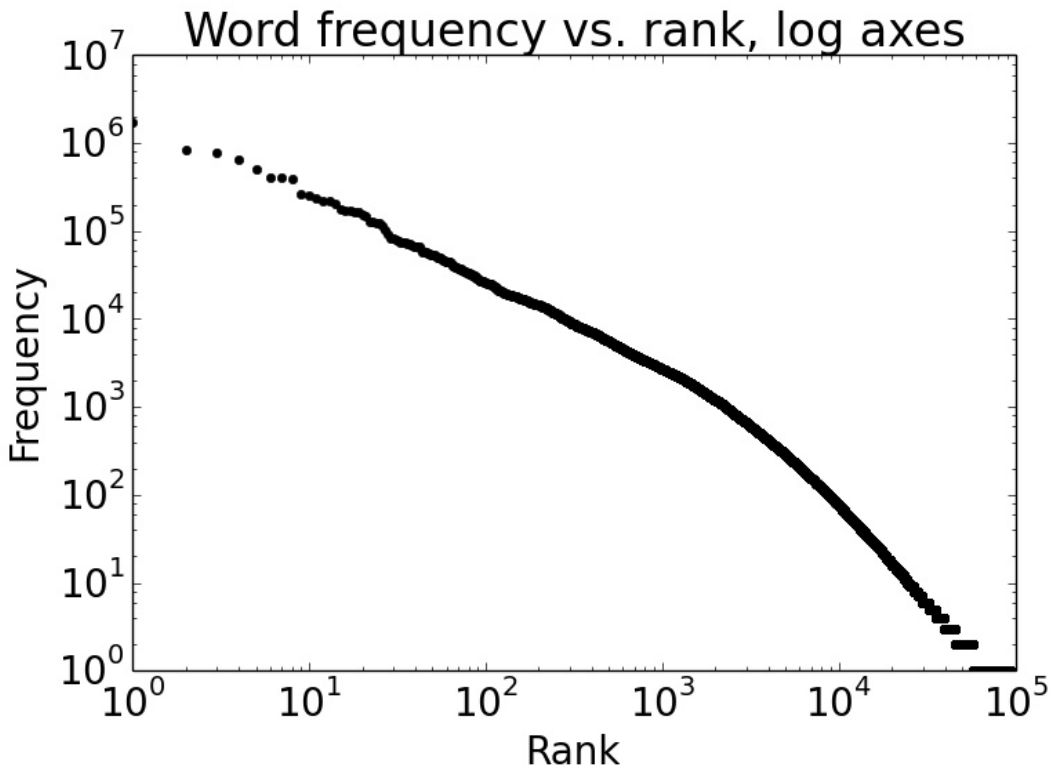
Order words by frequency. What is the freq of *n*th ranked word?



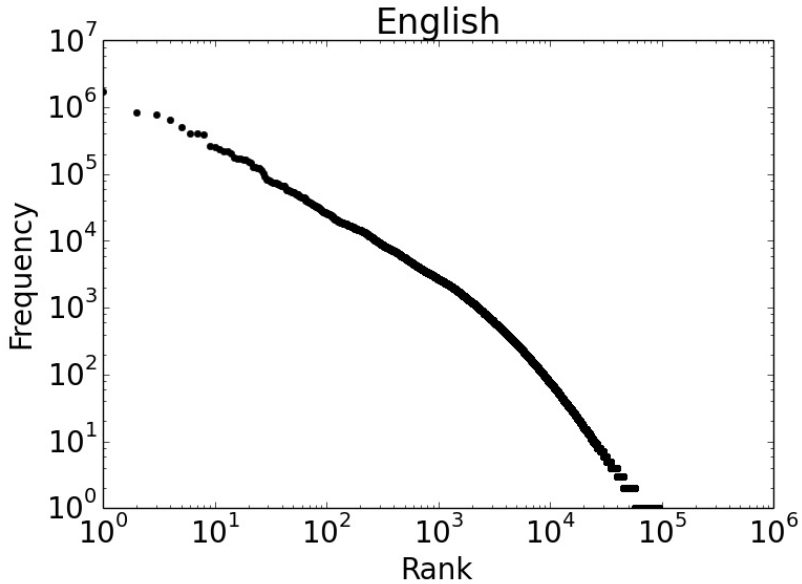
# Rescaling the axes

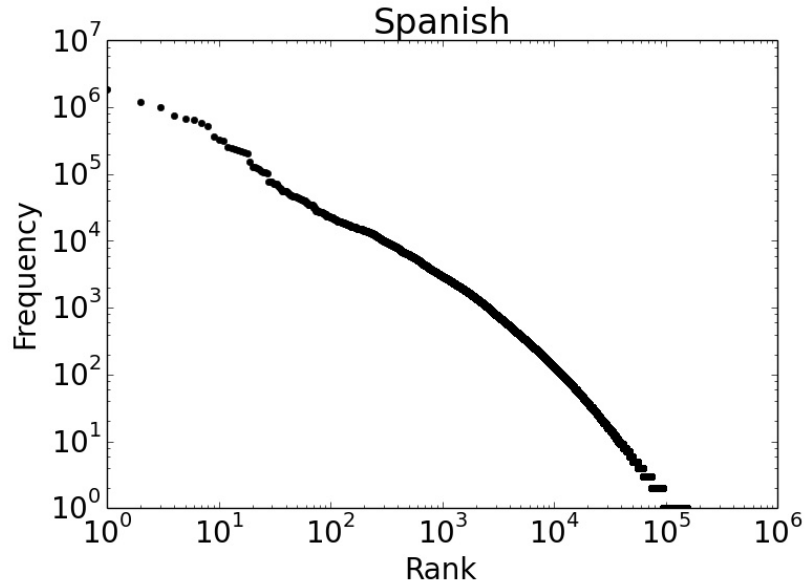
To really see what's going on, use logarithmic axes:

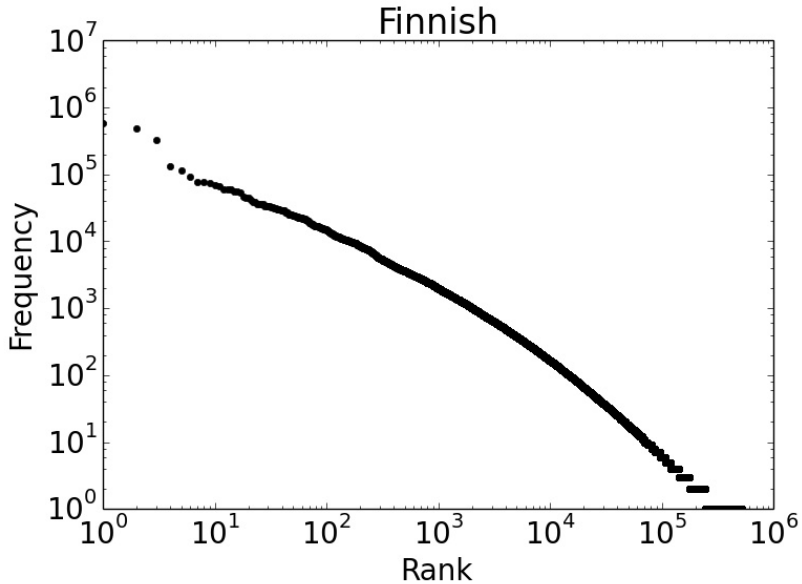


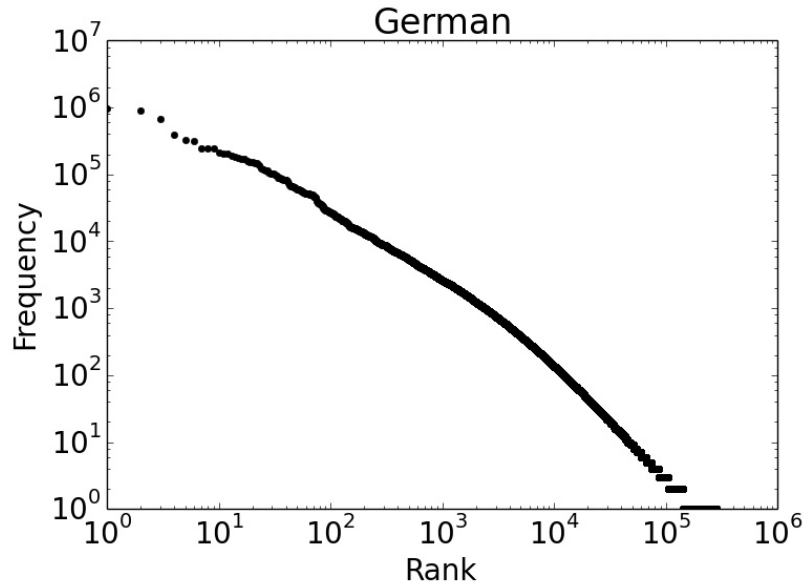

We will use logarithms again in this course. Please brush up on them if you are rusty.

# What about other languages?



# Zipf's law

Summarizes the behaviour we just saw:

$$f \times r \approx k$$

- $f$ = frequency of a word
- $r$ = rank of a word (if sorted by frequency)
- $k$ = a constant

# Zipf's law

Summarizes the behaviour we just saw:

$$f \times r \approx k$$

- $f$ = frequency of a word
- $r$ = rank of a word (if sorted by frequency)
- $k$ = a constant

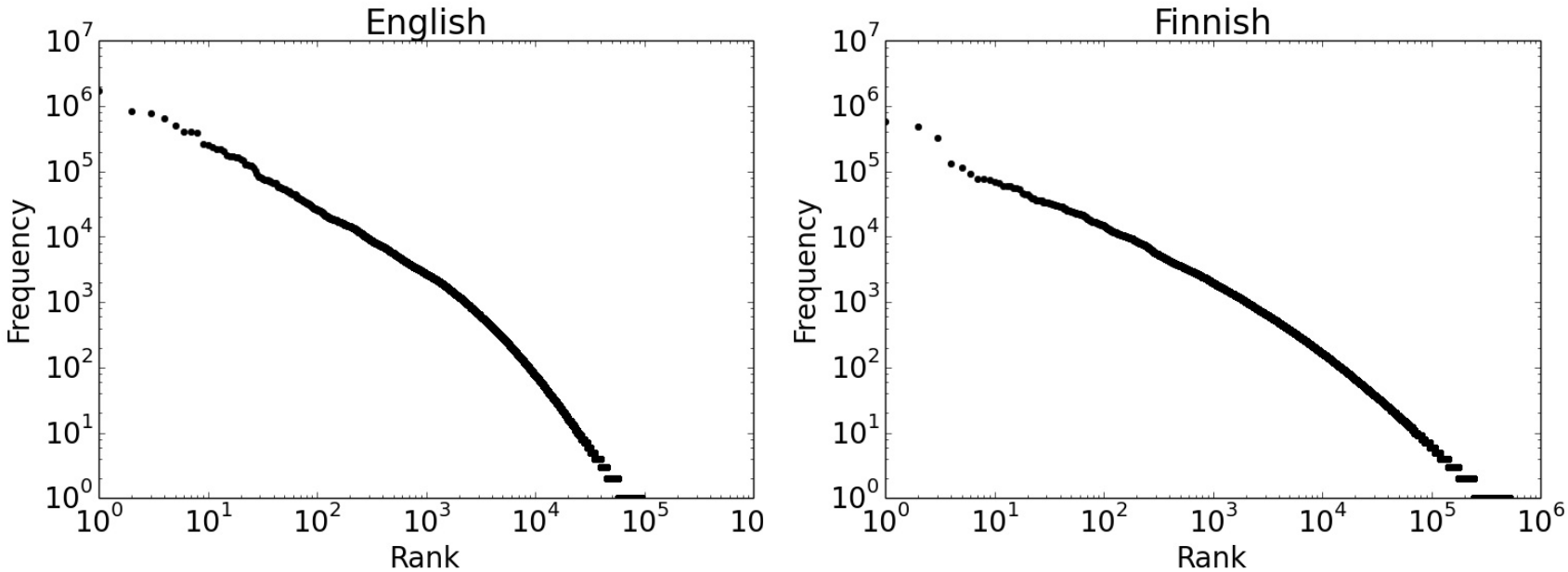
Why does Zipf's law look like a line in log-scale?

$$fr = k \quad \Rightarrow \quad f = \frac{k}{r} \quad \Rightarrow \quad \begin{aligned} \log f &= \log k - \log r \\ y &= c - x \end{aligned}$$

# Linguistics and Data

- Data
  - looking at real use of language in text
  - can learn a lot from empirical evidence
  - but: Zipf's law means there will always be rare instances
- Linguistics
  - build a better understanding of language structure
  - linguistic analysis points to what is important
  - but: many ambiguities cannot be explained easily

# Two plots in more detail




Although the shape is similar, the scale at the $x$-axis is different! What explains this?

# How Many Different Words?

10,000 sentences from the Europarl corpus

| Language | Different words |
|---|---|
| English | 16k |
| French | 22k |
| Dutch | 24k |
| Italian | 25k |
| Portuguese | 26k |
| Spanish | 26k |
| Danish | 29k |
| Swedish | 30k |
| German | 32k |
| Greek | 33k |
| Finnish | 55k |

Why the difference? **Morphology**: topic of next lecture.