Intelligent Data Analysis 2020

Lecture 2
Statistical Analysis of Texts

Martin Russell

UNIVERSITY OF BIRMINGHAM

Objectives

- Understand different approaches to text-based IR
  - Explicit knowledge vs data
- "Bundles of words" approaches
- Introduction to zipf.c
- Statistical analysis of word occurrence in text
- Zipf's Law
- Examples

Example Text

"There was no possibility of taking a walk that day. We had been wandering, indeed, in the leafless shrubbery an hour in the morning; but since dinner (Mrs. Reed, when there was no company, dined early) the cold winter wind had brought with it clouds so sombre, and a rain so penetrating, that further out-door exercise was now out of the question."

keywords

Charlotte Brontë, "Jane Eyre", first paragraph

Jane Eyre extract

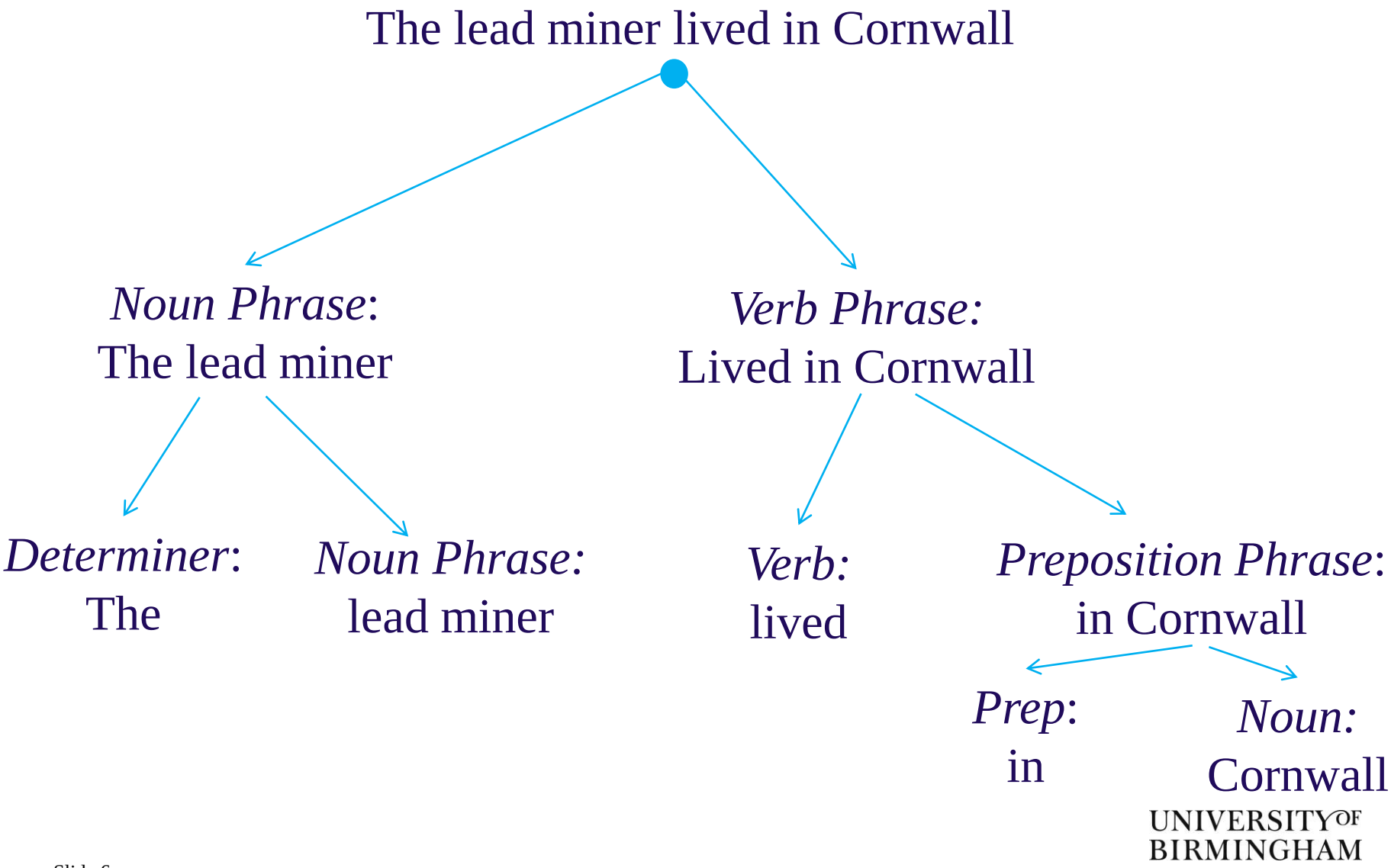
- What is it **about**?
- How do you know?
- What is your 'strategy' for understanding what a text is **about**?
- What are the component topics?
  - Exercise (walk, wandering, exercise)
  - Gardens (shrubbery)
  - Weather (cold, winter, wind, clouds, rain)

Structure in text

- Words
  - **Keywords** (some words are more important than others)
  - *Cold*, *Walk* and *Shrubbery* are important
  - *There*, *and* and *that* are not
- Sentences (Grammar / Syntax)
  - Word sequence structure helps us to understand and to remove ambiguity
  - 'Parts of speech'
    - *The lead miner lived in Cornwall*
    - *Keep that dog on a lead!*
    - *He won the lead role in the new film*

roles

Example



UNIVERSITY OF BIRMINGHAM

Knowledge vs. Data (1)

- Knowledge (Rationalism):
  - Try to copy human language processing (emulation)
- Two questions:
  - Do we understand sufficiently well how we do it?
  - Is our knowledge 'computationally useful'? I.e. is our knowledge sufficiently 'solid' to support algorithms and computer programs?
- These are topics in Natural Language Processing (NLP) and Computational Linguistics

Available knowledge

- Word inventories
  - Electronic dictionaries
- Word forms (noun, verb etc)
  - Available in electronic dictionaries
- Word meanings
  - Expressed in terms of predicate logic (properties)
- Grammar / syntax
  - Grammatical rules
- Parsers
  - Apply grammatical rules to a word sequence to determine if it is grammatical and, if so, its grammatical structure

Natural Language Processing

- Use word sense and meaning plus grammatical structure to infer 'meaning'
- Several problems
  - Grammar may be too accommodating – accept non-grammatical sentences
  - Grammar may be too restrictive – reject valid sentences
  - The number of interpretations of a simple sentence may be huge ("I saw the man on the hill with the telescope")
- Language is dynamic and changing

Knowledge vs. Data (2)

- Data (Empiricism, "Big Data")
  - Use **large** corpora of text instead of human knowledge
  - Use **machine-learning** to identify important structure and relationships
  - **Quantify** the problem
  - Rely on **quantities which can be measured** from these large corpora, rather than human opinion
- For example:
  - For each word *w* define a number *U*(*w*) which indicates how **useful** *w* is for Information Retrieval
  - Invent **algorithms** to find the **most useful** words
  - Invent **measures** of the **similarity** between queries and texts

Knowledge vs Data

- Need sophisticated computationally useful models of language and semantics to infer meaning
- Rational approaches accommodate complex structure but may be fragile and hard to generalise
  - "She ran, waving her hand in the air, across the bridge"
- Machine Learning (ML) is conceptually simpler, models are potentially huge, trained automatically
- NLP currently outperformed in most applications by ML – "Deep Learning", "Deep Neural Networks" (e.g: Amazon Echo/Alexa, etc)
- **Bundles of Words** approach to language processing

‘Bundles of Words’

There was no possibility of taking a walk that day. We had been wandering, indeed, in the leafless shrubbery an hour in the morning; but since dinner (Mrs. Reed, when there was no company, dined early) the cold winter wind had brought with it clouds so sombre, and a rain so penetrating, that further out-door exercise was now out of the question

the 4	early 1	walk 1
was 3	exercise 1	wandering 1
a 2	further 1	we 1
had 2	hour 1	when 1
in 2	indeed 1	wind 1
no 2	it 1	winter 1
of 2	leafless 1	with 1
so 2	morning 1	
that 2	mrs 1	
there 2	now 1	
an 1	out 1	
and 1	out-door 1	
been 1	penetrating 1	
brought 1	possibility 1	
but 1	question 1	
clouds 1	rain 1	
cold 1	reed 1	
company 1	shrubbery 1	
day 1	since 1	
dined 1	sombre 1	
dinner 1	taking 1	

What is a word?

- Tokens ≡ things separated by white space
- Hyphenation
  - Database ≡ Data-base?
- Case
  - “the bath shop” vs “the Bath shop”
  - “the brown house” vs “the Brown house”
- Morphology
  - retrieval, retrieve, retrieved, retrieving,…
- Punctuation
  - The ‘honest’ politician vs the honest politician

Some arbitrary choices…

- Tokens ≡ things separated by white space
- Ignore case:
  - London ≡ london
  - BBC ≡ bbc
- Ignore non-alphanumerics at start and end of token:
  - ‘honest’ ≡ honest. ≡ honest! ≡ “honest ≡ honest

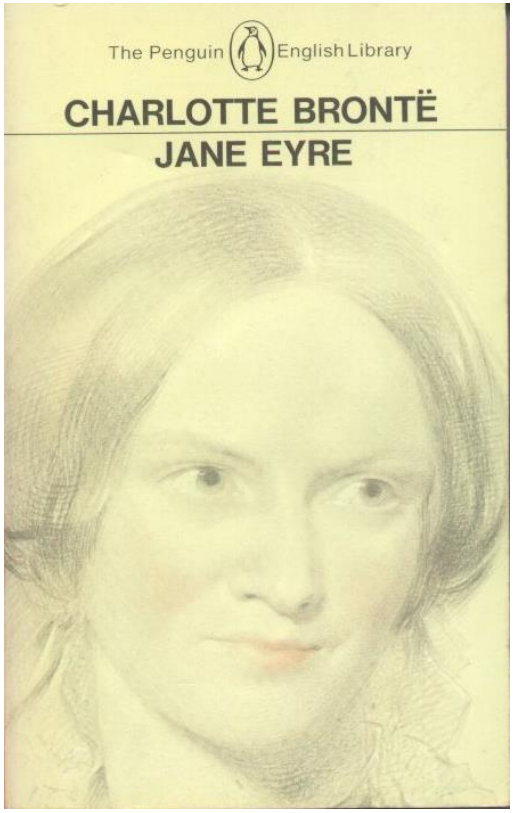
Analysis of Word Frequency

- `zipf.c`
  - ANSII C program for simple analysis of texts
  - Finds the set of different tokens in the text
  - Counts how many times each word occurs
  - Orders words according to the number of times they occur in the text (their **rank**)
  - Prints out the result, and
  - Stores results in a file `results`

Compilation of “Data Mining” C code

- Simple ANSII C
- OS independent – should work on any platform with any ANSII-compliant C compiler
- Download from course website
- Compile using MS Visual Studio .NET **command line**
- `cl zipf.c`

Statistical Analysis of Frequency



- Complete novels available online: http:/www.literature.org
- Start with "Jane Eyre", Charlotte Brontë, 1847
- Penguin Edition - 489 pages
- 1,039 KBytes

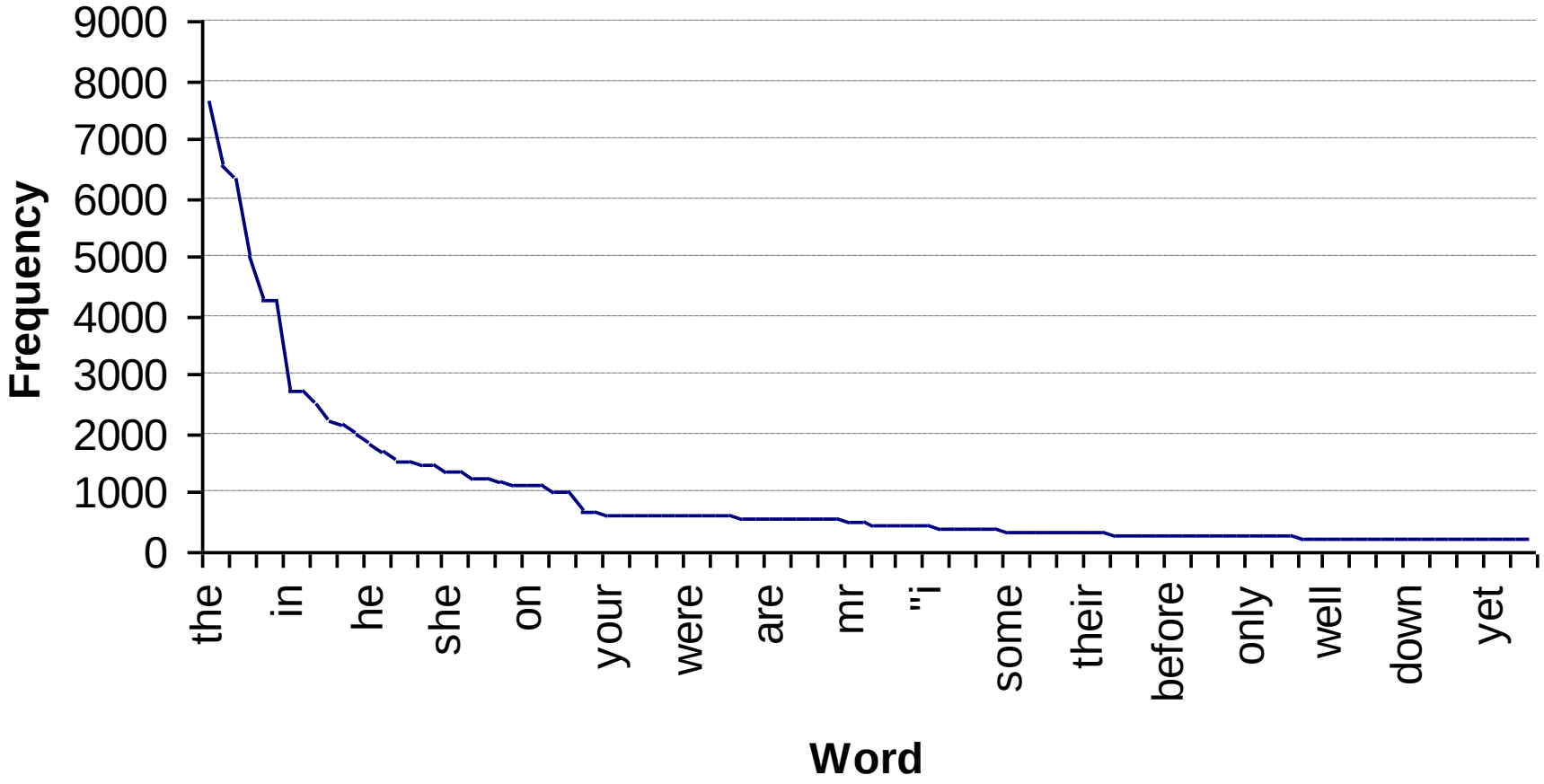
Top 10 words in Jane Eyre

Top 10		101-110		7861-7870	
the	7638	can	218	abate	1
i	6536	about	217	abbot's	1
and	6335	looked	216	abigail	1
to	5028	think	213	abilities	1
of	4299	seemed	209	abode--whether	1
a	4294	day	206	abodes	1
in	2717	any	204	abominable	1
you	2709	own	203	abrid	1
was	2495	much	200	abruptness	1
it	2219	come	199	absences	1

Different words 15,827, Total words 184,640

UNIVERSITY OF BIRMINGHAM

Word frequency plot for Jane Eyre



Zipf's Law

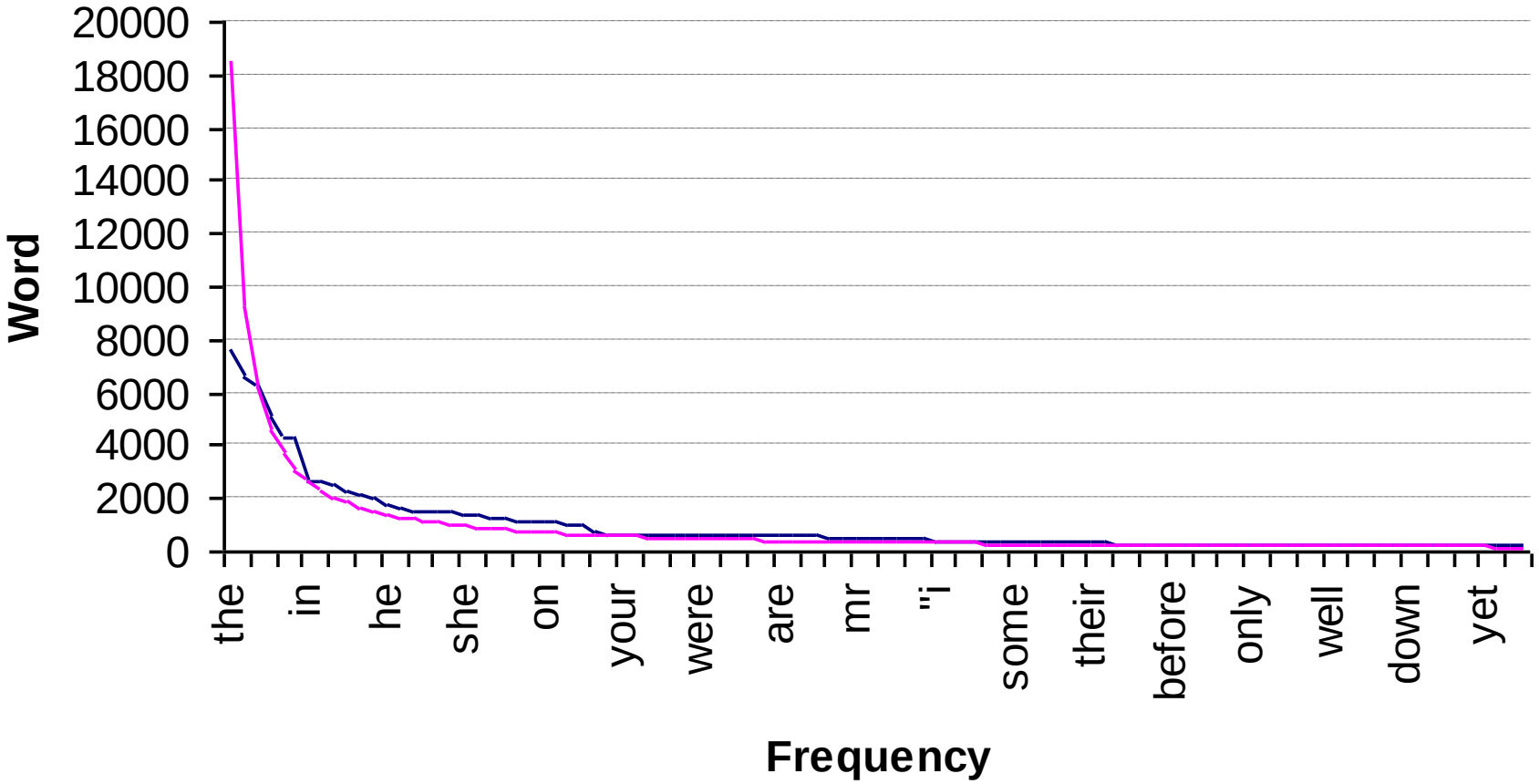
- George Kingsley Zipf (1902-1950)
  - For each word *w*, let *F*(*w*) be the number of times *w* occurs in the corpus
  - Sort the words according to frequency
  - The word's rank-frequency distribution will be fitted closely by the function:

$$F(r) = \frac{C}{r^{\alpha}}, \text{ where } \alpha \approx 1, C \approx 0.1$$

Zipf's Law

total = 184,640
vocal = 15,824
% < 10 occurs?
$f(w) = \frac{0.1}{r}$ $f(r) = \frac{10}{184640} = \frac{0.1}{r}$ $r = 18466.4$
$\frac{13990}{15824} \approx 88\%$

Zipf's law —— Actual statistics from "Jane Eyre" ——



UNIVERSITY OF BIRMINGHAM

Zipf's Law (logarithm form)

$$F(r) = \frac{C}{r^{\alpha}}, \text{ where } \alpha \approx 1, C \approx 0.1$$

Therefore,

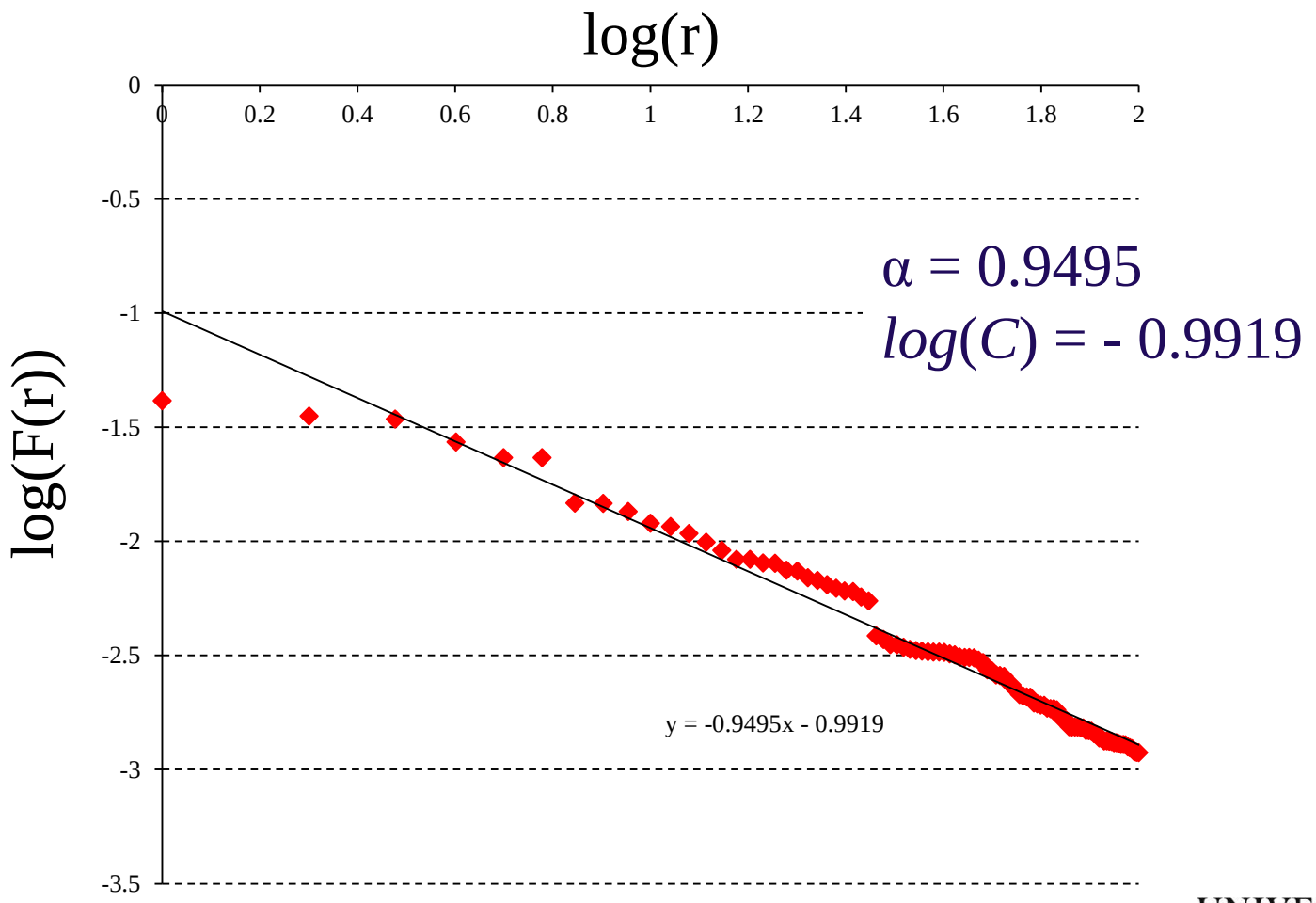
$$\log(F(r)) = \log(C) - \alpha \log(r)$$

$\log(F(r))$

- On a log-log scale, Zipf's Law predicts a straight-line relationship between log-rank and log-frequency, where α is the slope of the line and *C* is the intersection with the vertical axis
- This provides a way to estimate *C* and *α*

Zipf's Law (logarithm form)

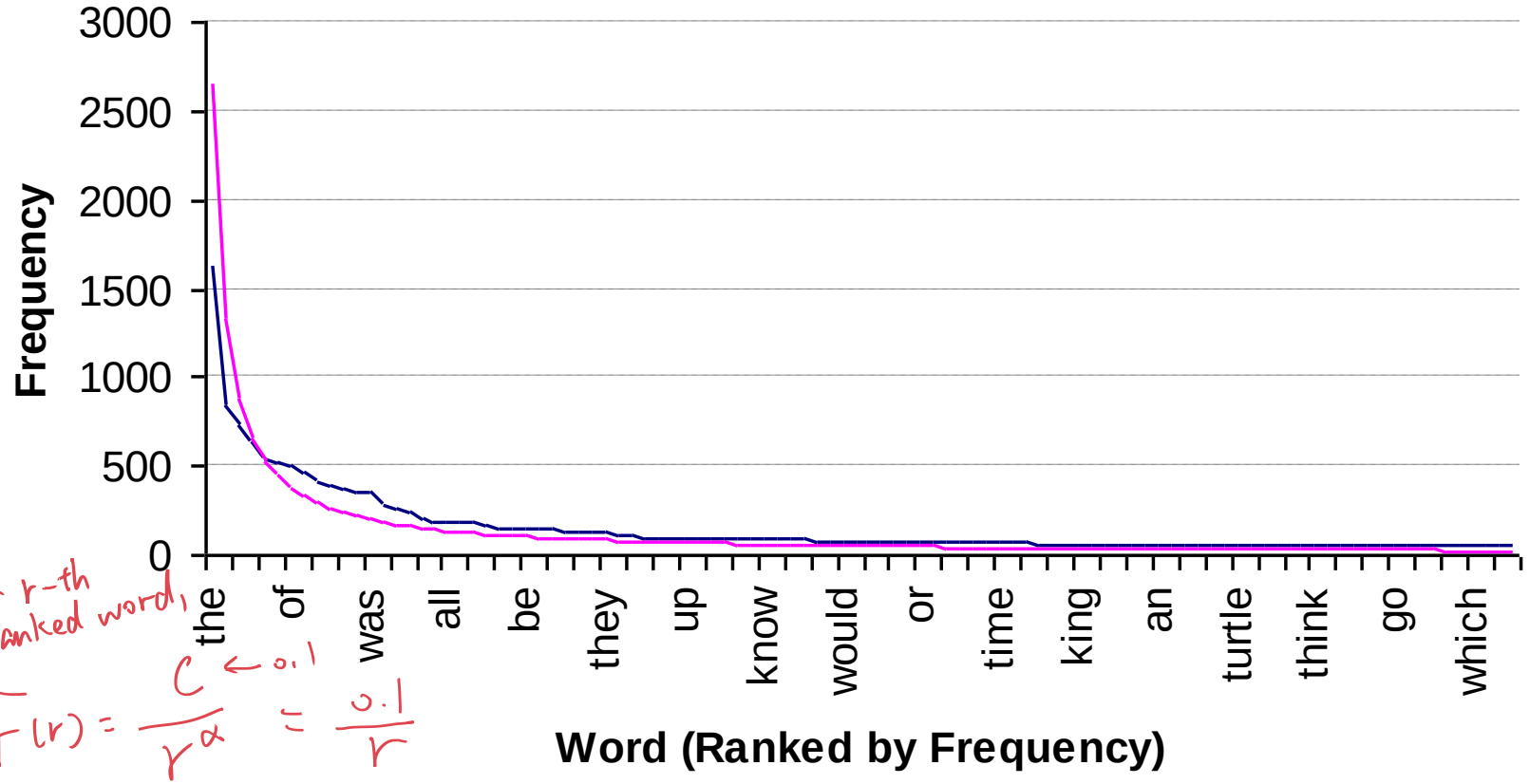
Zipf's Law ——— Actual statistics from "Jane Eyre" ◆



UNIVERSITY OF BIRMINGHAM

Word Frequency Plot: Alice in Wonderland

Zipf's law — Actual statistics from "Alice in Wonderland"

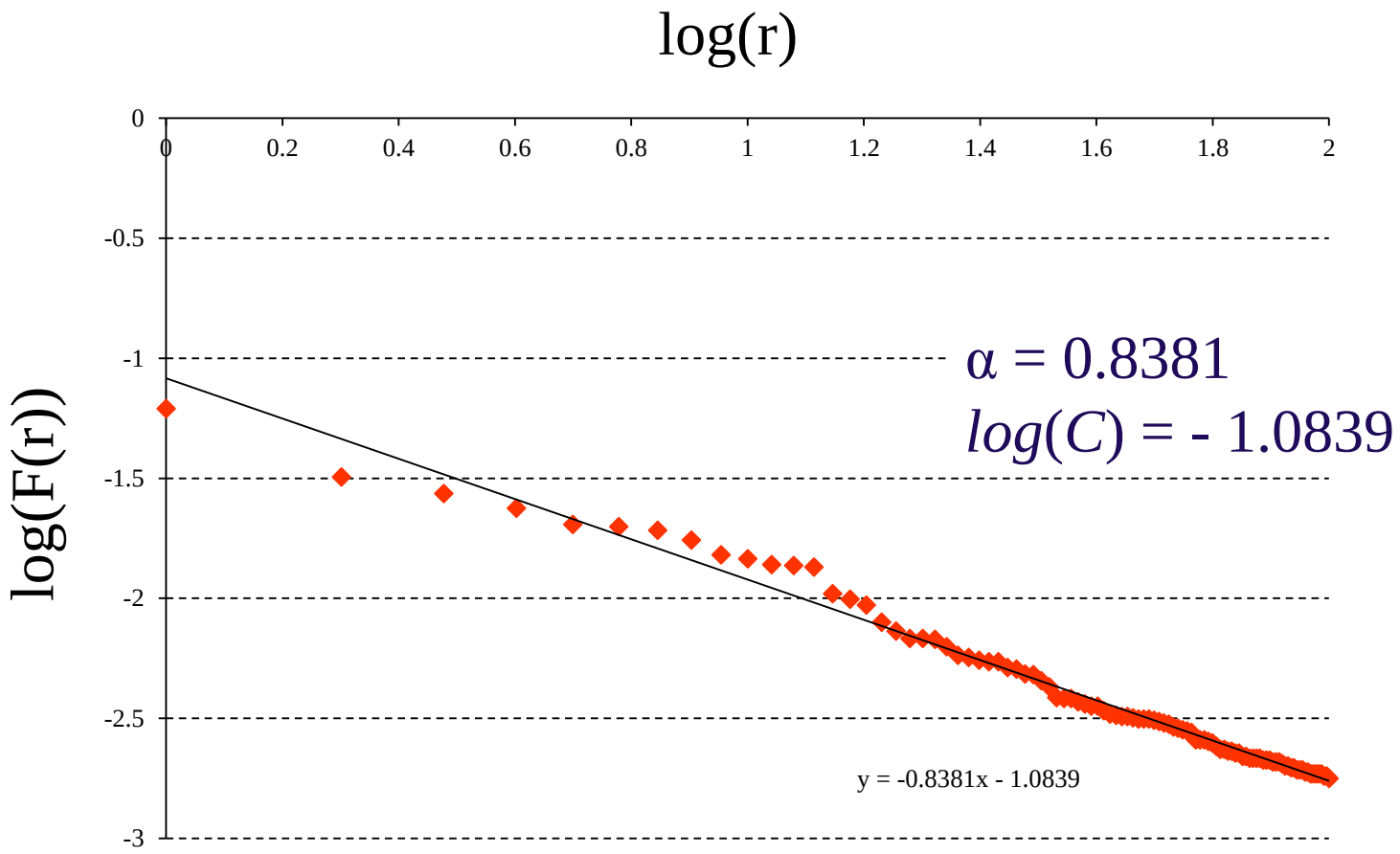


for r-th ranked word, $F(r) = \frac{C}{r^{\alpha}} = \frac{0.1}{r}$ ← 0.1

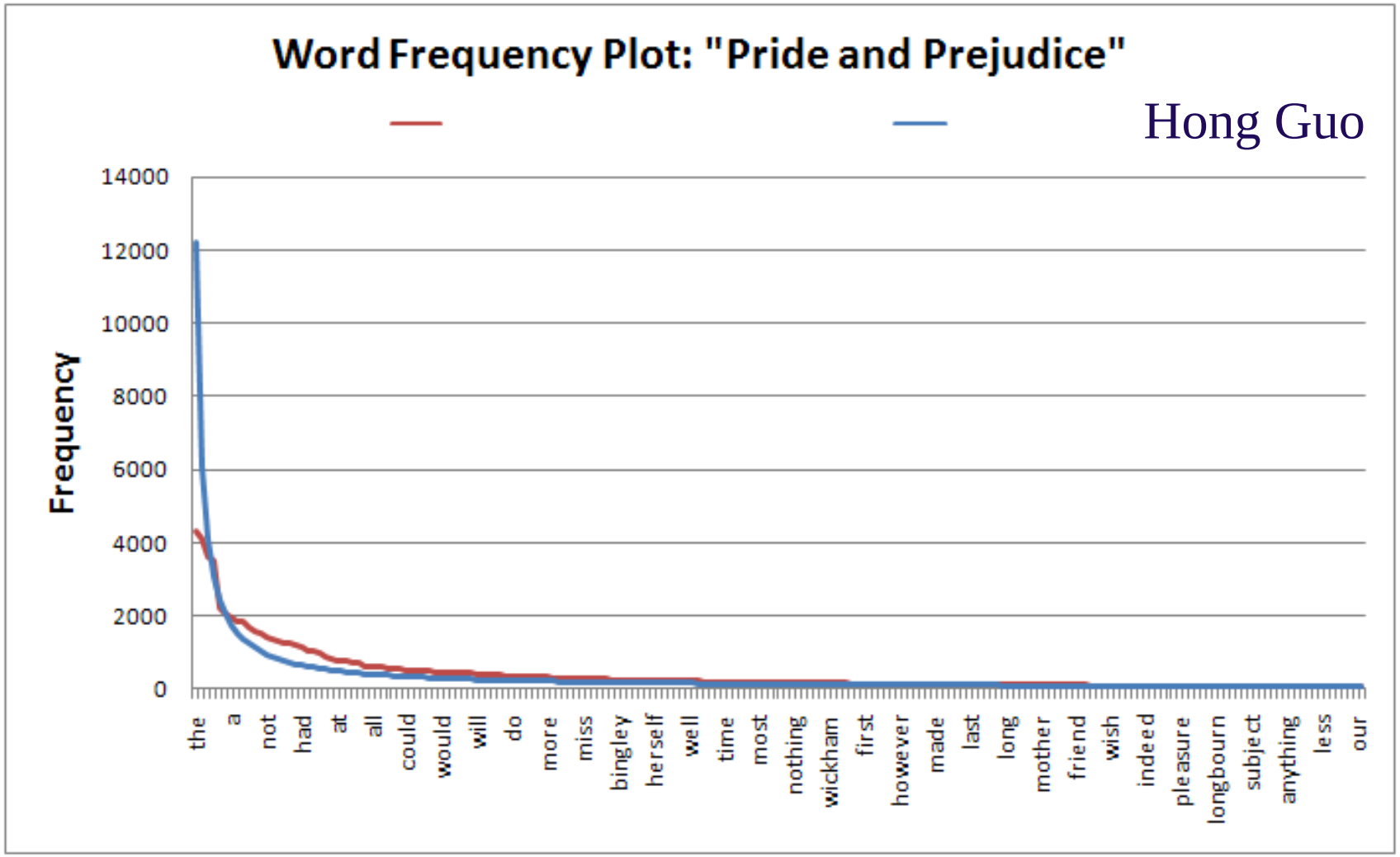
Different words 2,787, Total words 26,395

$F(r) = \frac{\text{no}}{\text{total}}$ $\frac{0.1}{r} = \frac{10}{26395}$ $r = 263.95 \approx 264$ per: $\frac{2787 - 264}{2787} = 90\%$

Log-log plot – Alice in Wonderland



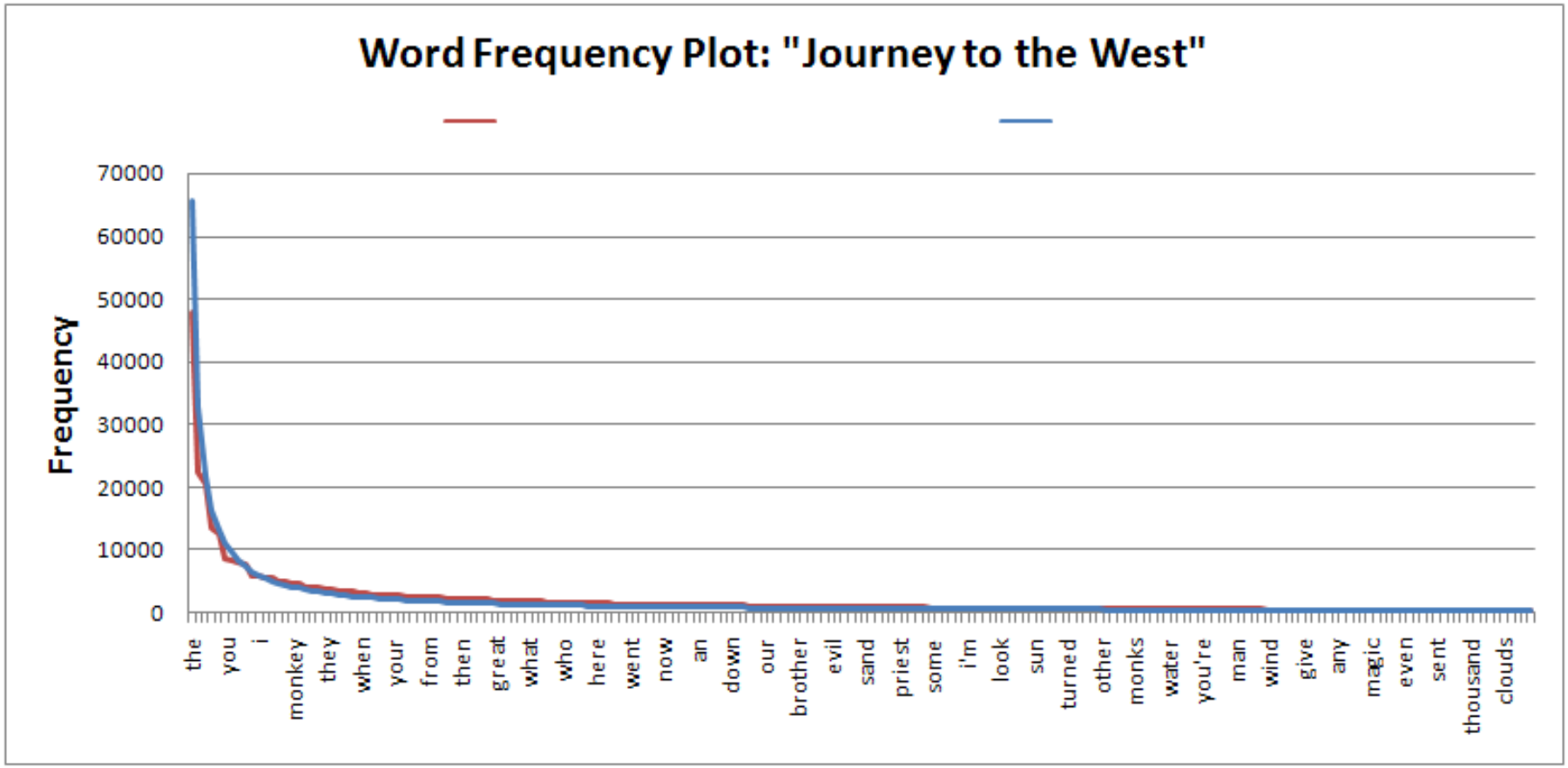
Zipf - Pride and Prejudice



UNIVERSITY OF BIRMINGHAM

Zipf vs “Journey to the West”

Hong Guo



UNIVERSITY OF BIRMINGHAM

Some non-text examples

- Mathematics Today, vol. 47, no. 5, October 2011
- "Urban maths – Zipf's Law"
  - Populations of the countries of the world
  - UK new car sales 2010
  - Counts of first digit from 1,836 equity prices quoted in The Times

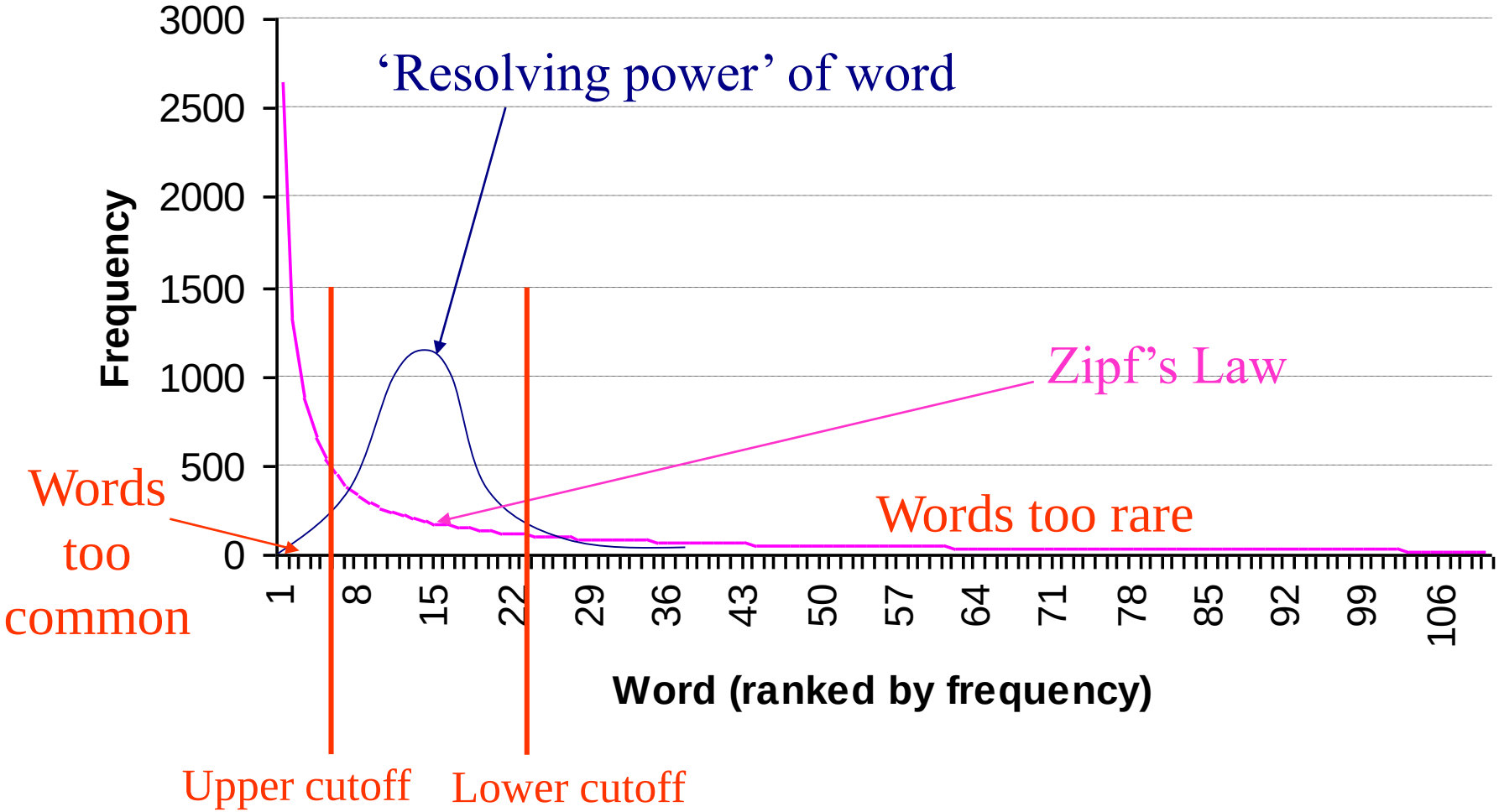
Zipf's Law

- Why does it hold?
- Is it relevant to Information Retrieval?

Why does Zipf's Law work?

- Zipf's law appears to reflect a number of factors:
  - The requirements of humans to communicate
    - Use as little effort as possible to successfully communicate a message
  - Basic combinatorics
  - The requirement of grammar for simple 'glue' words
  - Author and topic vocabularies

‘Resolving Power’ of words



Homework

- Calculate α and C for the PayPal UserAgreement
  - Download the PayPal user agreement
  - Download zipf.c from the course website
  - Compile it under your favourite OS (see hints in comments at top of source)
  - Plot the result on a log-log scale using Excel
  - Find the best straight-line fit