Corpus Linguistics: Analysing word frequencies

1

INTRODUCTION

A corpus of British English and American English

WORD FREQUENCIES

- Books from 19th century British and American writers downloaded from the Gutenberg Project
- Number of individual words: 30 723
- Number of occurrences in AE corpus: 11 709 009
- Number of occurrences in BE corpus: 13 795 791
- Total size of corpus: 25 504 800

WORD FREQUENCIES

Is the word "colour" used more often in American or British English?

	Occ. of "colour"	Total number of words	Frequency
AE	255	11 709 009	21.77 pwm
BE	1772	13 795 791	128.44 pmw

Is it significant ?

CHI SQUARE

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

WORD FREQUENCIES

CHI SQUARE

$$\chi^2 = 906.71$$

$$p < .001$$

WORD FREQUENCIES

Is the word "the" used more often in American or British English?

	Occ. of "the"	Total number of words	Frequency
AE	848 729	11 709 009	72 485 pwm
BE	914 669	13 795 791	66 300 pmw

WORD FREQUENCIES

CHI SQUARE

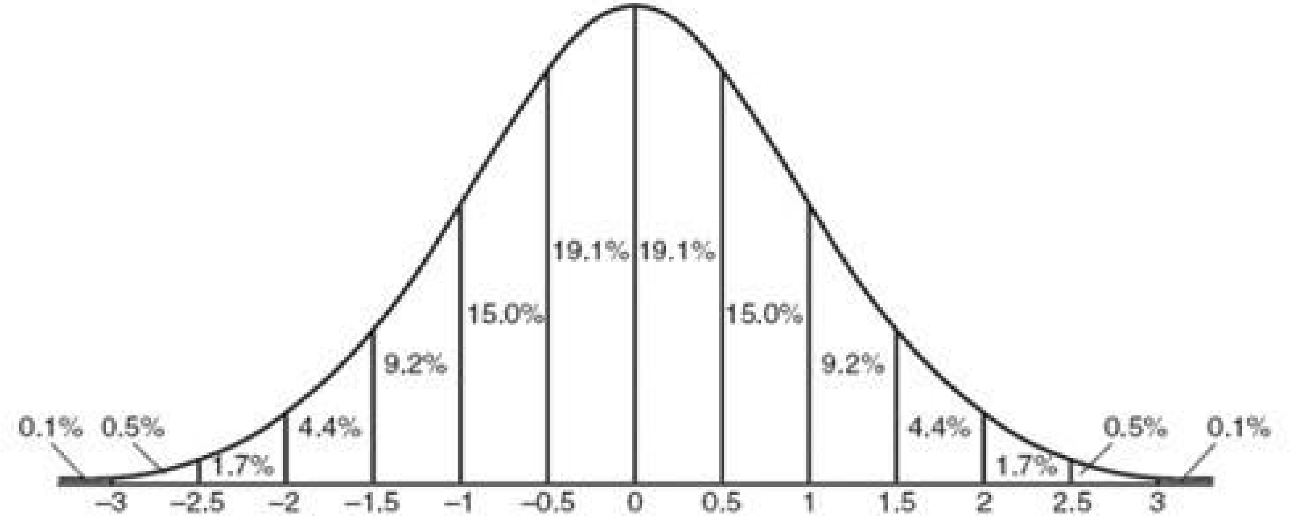
$$\chi^2 = 3503.73$$

$$p < .001$$

But...

Is χ^2 really appropriate?

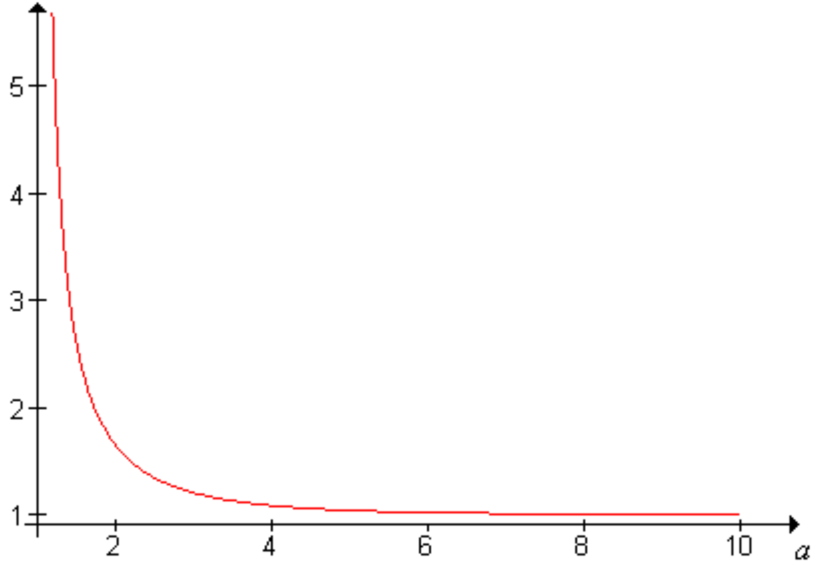
Normal distribution: 95% of the values lie within two standard deviations



Word frequencies: zeta distribution

ZIPF’S LAW

A small number of words have a very high frequency, and a large number of words have a very low frequency.



NULL HYPOTHESIS

The null hypothesis is the idea that there is no relationship between two measured phenomena.

IN OTHER WORDS…

The null hypothesis is the hypothesis that chance alone can explain what we're observing.

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT *YEARS* AGO.

LANGUAGE IS *NOT* RANDOM

> *“Words are not selected at random. There is no a priori reason to expect them to behave as if they had been, and indeed they do not.”*
>
> Adam Kilgarriff, 1996

1

Chi-square:

How bad is it?

USING CHI-SQUARE ON THE WHOLE CORPUS:

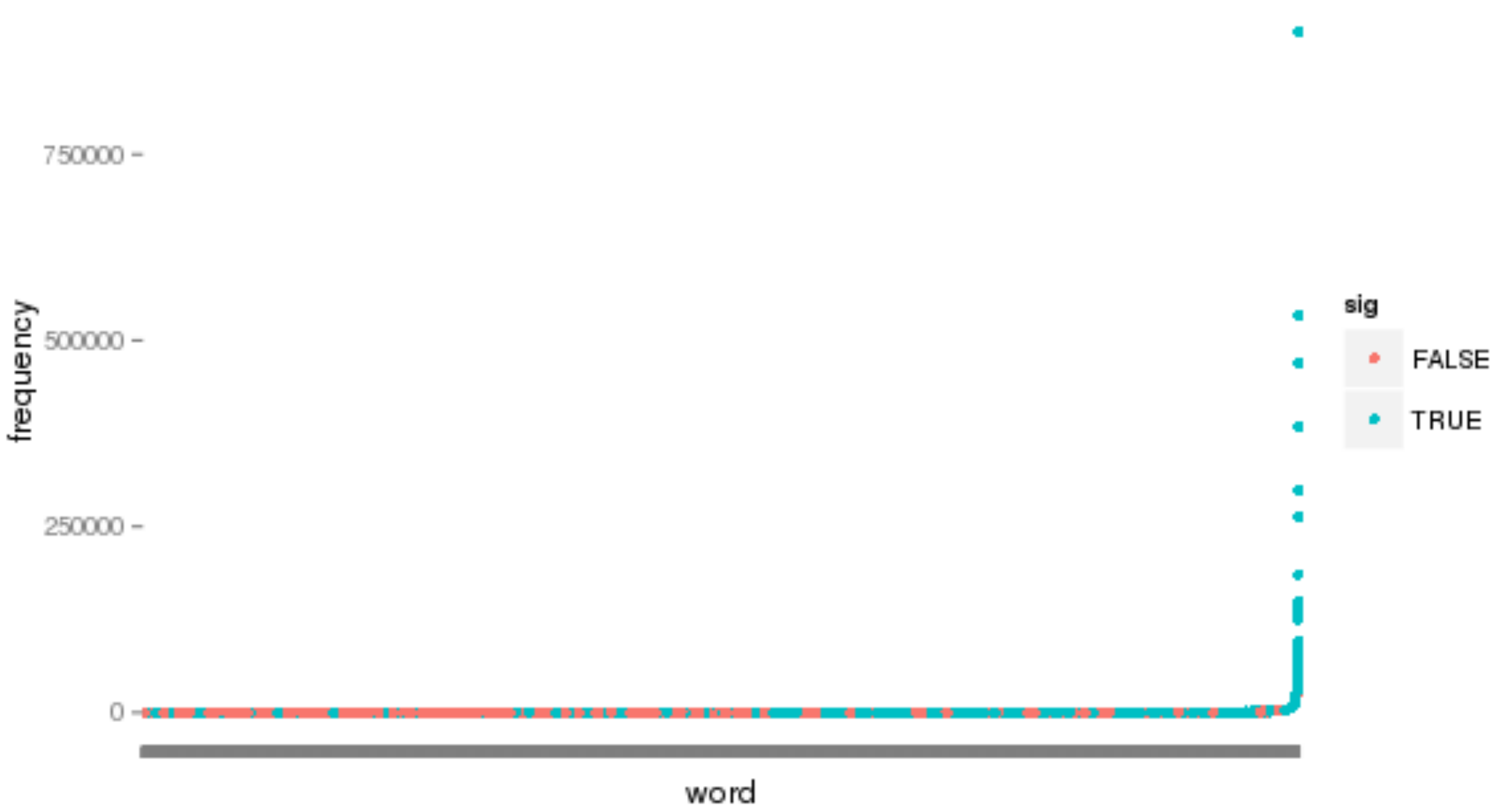
- Only keep words with over 5 occurrences
- Only keep words that occur in both AE and BE corpus

USING CHI-SQUARE ON THE WHOLE CORPUS:

- ❖ Only keep words with over 5 occurrences
- ❖ Only keep words that occur in both AE and BE corpus

- ❖ Number of significant results: **15197**
- ❖ Number of non-significant results: **15526**
- ❖ **49.4%** of tests turn out significant (p < .05)

WORD FREQUENCIES



2

ALTERNATIVE #1 :

Cramer’s V

Cramer’s Phi & Cramer’s V

$$\Phi = \sqrt{\frac{\chi^2}{N}}$$

$$V = \sqrt{\frac{\Phi}{k-1}}$$

USING CRAMER’S PHI ON THE WHOLE CORPUS:

- Same data as for Chi-square test
- Maximum value of Phi coefficient is determined by the distribution of the two variables

USING CRAMER'S PHI ON THE WHOLE CORPUS:

- ❖ Same data as for Chi-square test
- ❖ Maximum value of Phi coefficient is determined by the distribution of the two variables

$\Phi > 0.5$:

- ❖ Significant results: **1198**
- ❖ Non-significant results: **29525**
- ❖ **4.06%** of tests turn out significant

WORD FREQUENCIES

USING CRAMER’S PHI ON THE WHOLE CORPUS:

- Same data as for Chi-square test
- Maximum value of Phi coefficient is determined by the distribution of the two variables

Φ > 0.5 :

- Significant results: **1198**
- Non-significant results: **29525**
- **4.06%** of tests turn out significant

Φ > 0.6 :

- Significant results: **698**
- Non-significant results: **30025**
- **2.32%** of tests turn out significant

WORD FREQUENCIES

USING CRAMER’S PHI ON THE WHOLE CORPUS:

- Phi coefficient for “colour”: **0.447**
- Phi coefficient for “the”: **0.001**

3

ALTERNATIVE #2 :

Wilcoxon-Mann-Whitney ranking test

Wilcoxon-Mann-Whitney ranking test

WMW

- Uses frequency to rank items and determine the value of the statistic (U)
- Divide the data in equal sized samples
- For each observation, retain frequency and origin of the sample (AE or BE)

“raining”					
1	2	4	4	5	7
AE	AE	BE	AE	BE	BE
1	2	3	4	5	6

WORD FREQUENCIES

WMW

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$$

R1 = sum of ranks in sample 1
n1 = sample size for sample 1

- The significance of the U statistic can be checked using normal distribution tables.
- AE and BE were divided in 10 equal sized chunks
- Tests made on all words with a frequency over 30 (n = 15756)

WORD FREQUENCIES

WMW

$p < 0.05$:

- Significant results: **2357**
- Non-significant results: **13399**
- **17.59%** of tests turn out significant

WORD FREQUENCIES

WMW

$p < 0.05$:

- Significant results: **2357**
- Non-significant results: **13399**
- **17.59%** of tests turn out significant

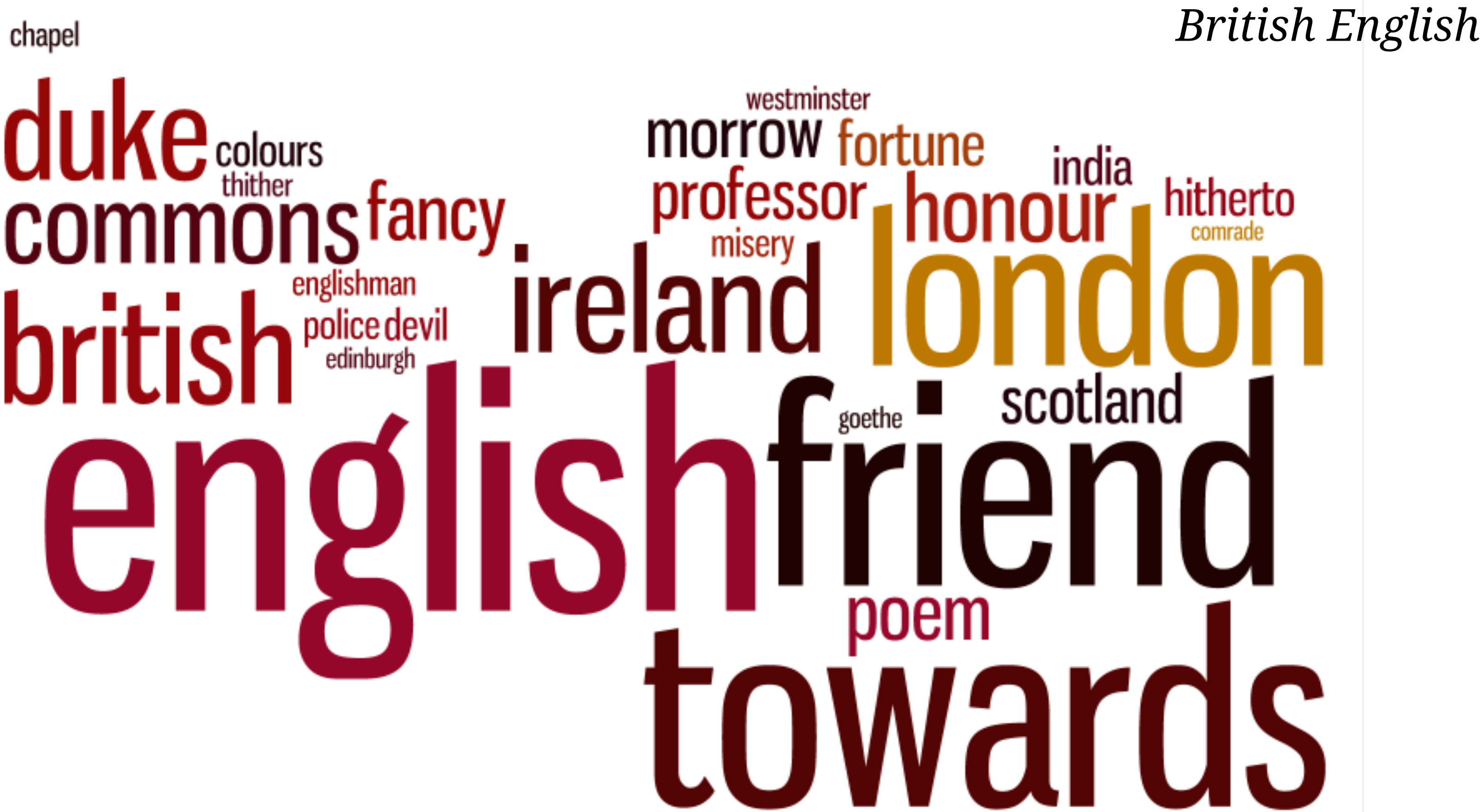
$p < 0.01$:

- Significant results: **889**
- Non-significant results: **14867**
- **5.98%** of tests turn out significant

WORD FREQUENCIES

Let’s see some results…

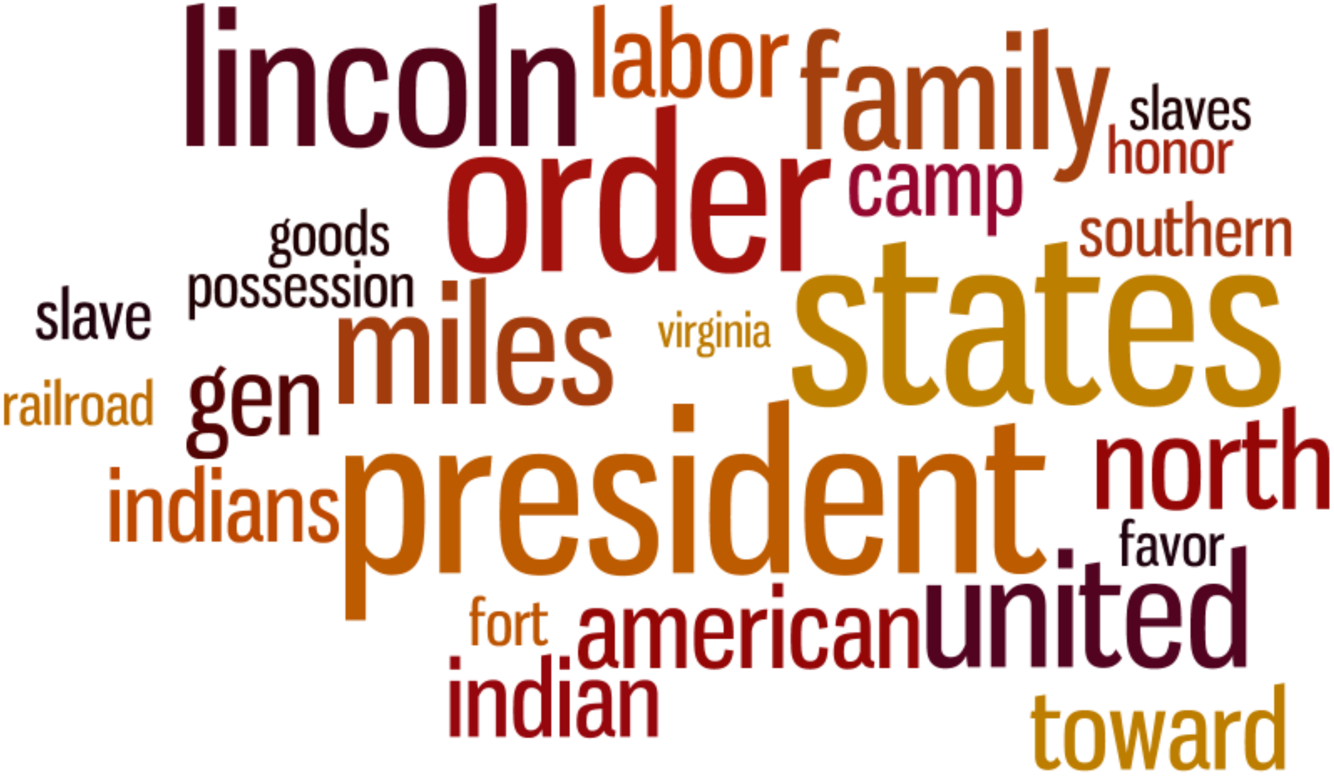
WORD FREQUENCIES



British English

WORD FREQUENCIES

American English



4

CONCLUSION

Choosing the most appropriate test

WORD FREQUENCIES

CRAMER’s V

- ☺ No local copy of corpus needed
- ☺ No programming skills required
- ☹ Interpretation can be difficult

WILCOXON-MANN-WHITNEY

- ☹ Local copy of corpus needed
- ☹ Some programming skills required
- ☺ Interpretation is easy

THANKS!

Any questions?