Taylor’s law for Human Linguistic Sequences

Tatsuru Kobayashi

Kumiko Tanaka-Ishii

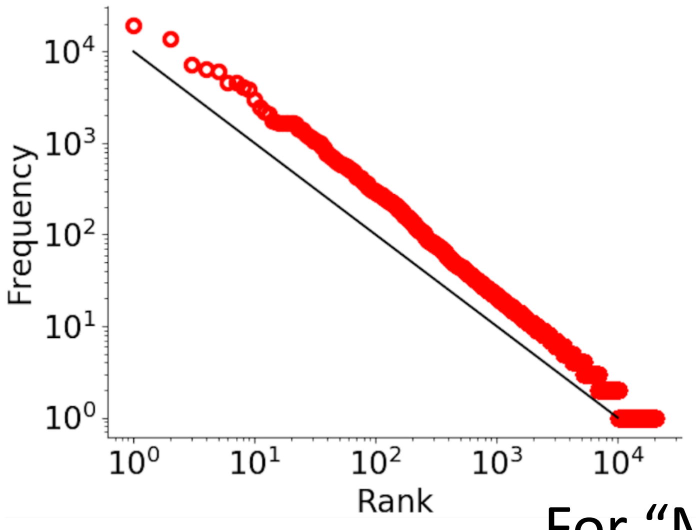
Research Center for Advanced Science Technology

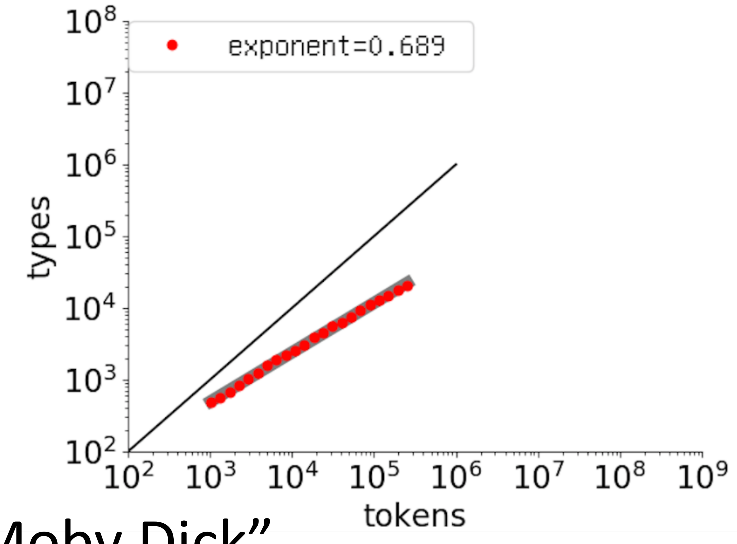
The University of Tokyo

Power laws of natural language

1. Vocabulary Population

- Zipf's law
- Heaps' law





For "Moby Dick"

2. Burstiness ⇐ About how the words are aligned

Words occur in clusters

Occurrences of words fluctuate

⇒ These can be analyzed through power laws

Today's talk is about quantifying the degree of fluctuation.
How these could be useful will be presented at the end.

Fluctuation underlying text

Any words (any word, any set of words) occur in clusters

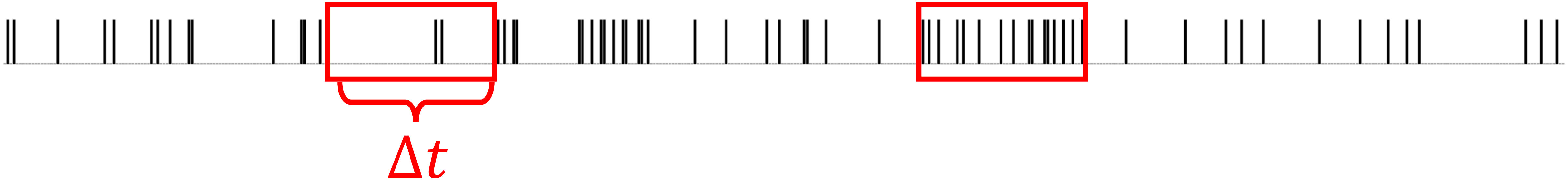
Occurrences of rare words in Moby Dick (below 3162th)

Two ways of analysis

- Fluctuation analysis
- Long range correlation → weaknesses

Fluctuation underlying text → Look at variance in Δt

Any words (any word, any set of words) occur in clusters
Occurrences of rare words in Moby Dick (below 3162th)



Variance is larger when events are clustered vs. random

Two ways of analysis

- Fluctuation analysis
- Long range correlation

- Fluctuation Analysis (Ebeling 1994)
  variance w.r.t. Δt
- Taylor's analysis ← Our achievements
  variance w.r.t. mean

Taylor's law (Smith, 1938; Taylor, 1961)

Power law between standard deviation and mean of event occurrences within (space or) time Δt

$$\sigma \propto \mu^{\alpha}$$

Empirically $0.5 \leq \alpha \leq 1.0$ (but $\alpha < 0.5$ is of course possible, too)

Empirically known to hold in vast fields (Eisler, 2007)

ecology, life science, physics, finance, human dynamics …

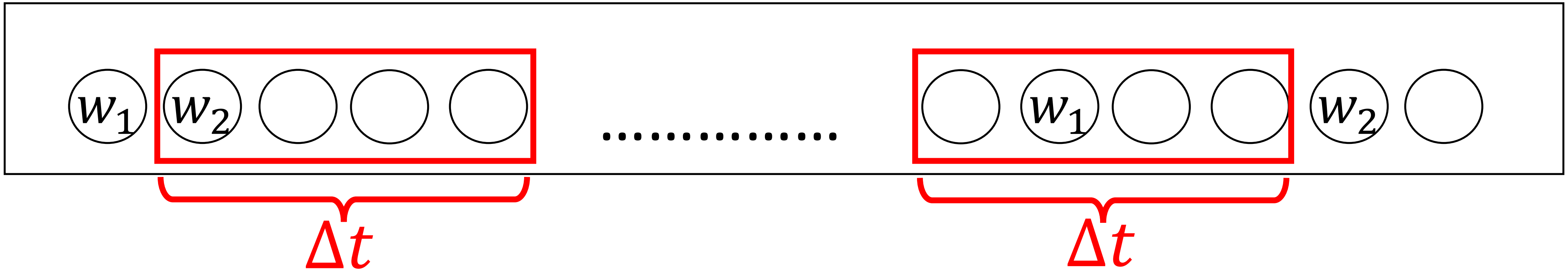
The only application to language is

Gerlach & Altmann (2014) ← not really Taylor analysis

We devised a new method based on the original concept of Taylor's law

Our method

Word sequence (text)



1 For every word kind $w_k \in W$ count its number of occurrence within given length Δt.

2 Obtain mean μ_k and standard deviation σ_k of w_k.

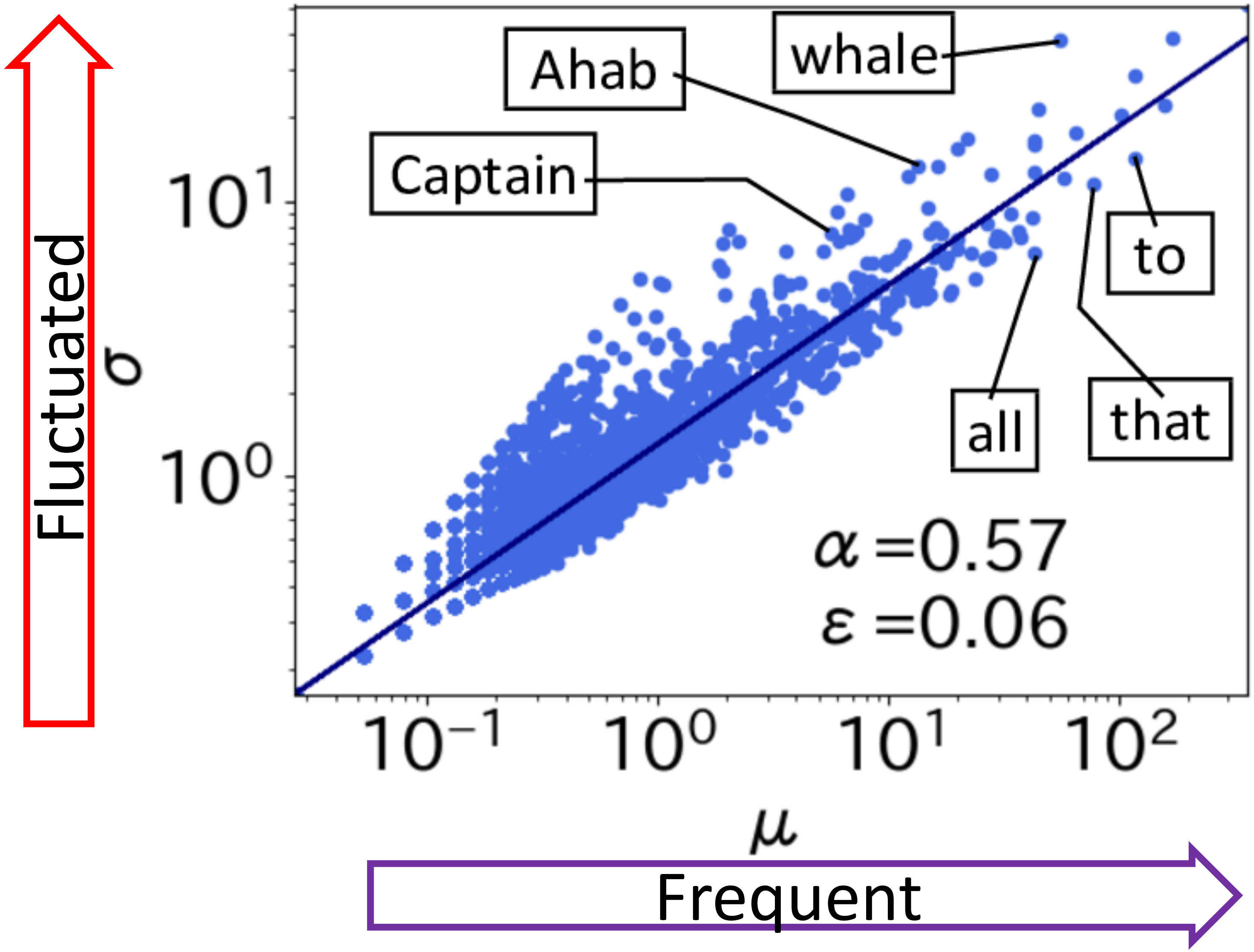
3 Plot μ_k and σ_k for all words.

4 Estimate α using the least squares method in log scale

$$\hat{c}, \hat{\alpha} = \mathrm{argmin}_{c,\alpha} \epsilon(c, \alpha),$$

$$\epsilon(c, \alpha) = \sqrt{\frac{1}{|W|} \sum_{k=1}^{|W|} (\log \sigma_k - \log c \mu_k^{\alpha})^2}.$$

Taylor's law of natural language



'Moby Dick'

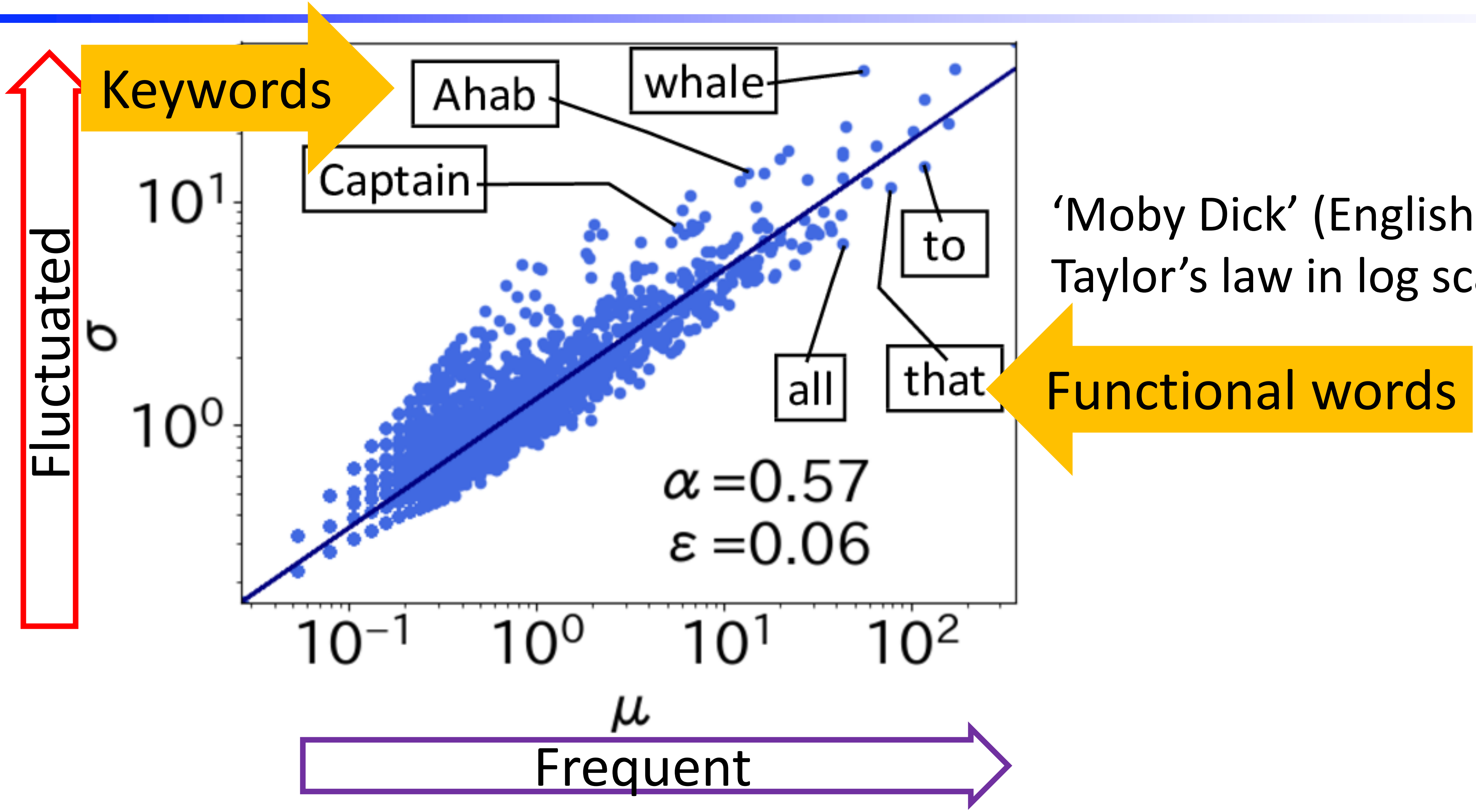
English, 250k words,

vocabulary size 20k words

Taylor's law in log scale

- Here, $\Delta t \approx 5000$.
- Every point is a word kind
- Estimated Taylor exponent $\alpha = 0.57$.
- Taylor exponent α corresponds to gradient of $\log \mu$-$\log \sigma$ plot.

Taylor's law of natural language



'Moby Dick' (English)'s Taylor's law in log scale

Theoretical analysis of the exponent

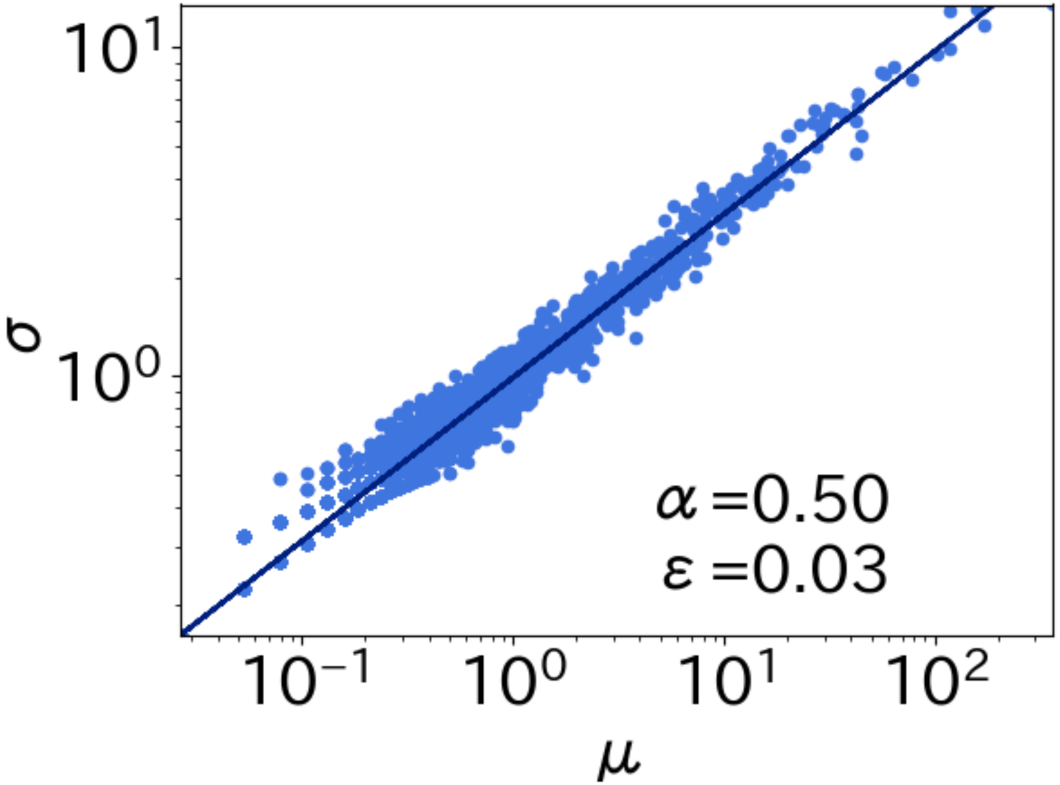
Empirically $0.5 \leq \alpha \leq 1.0$

$\alpha = 0.5$

if all words are independent and identically distributed (i.i.d.).

Shuffled ‘Moby Dick’
$\Delta t \approx 5000.$

Taylor Exponent $\alpha = 0.5$
because shuffled text is
equivalent to i.i.d. process.



Theoretical analysis of the exponent

$\alpha = 1.0$

if words always co-occur with the same proportion.

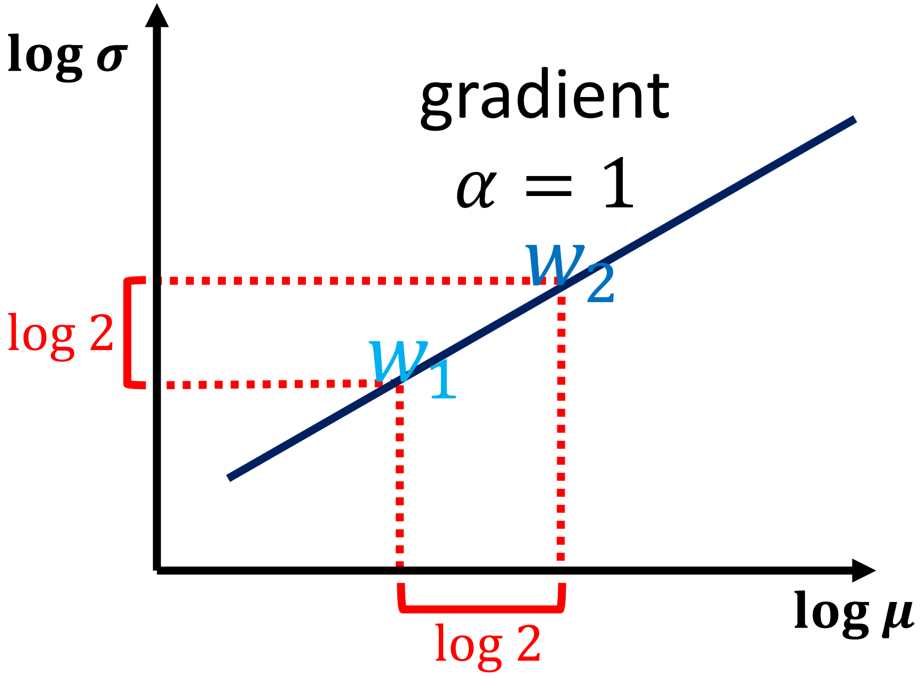
ex) Suppose that $W = \{w_1, w_2\}$, and w_2 occurs always twice as w_1

… $w_1: 17, w_2: 34$ … $w_1: 3, w_2: 6$ …

Δt

$\Longrightarrow \mu_2 = 2\mu_1, \sigma_2 = 2\sigma_1$

$\Longrightarrow \sigma \propto \mu$



Taylor's law for other data

Child directed speech

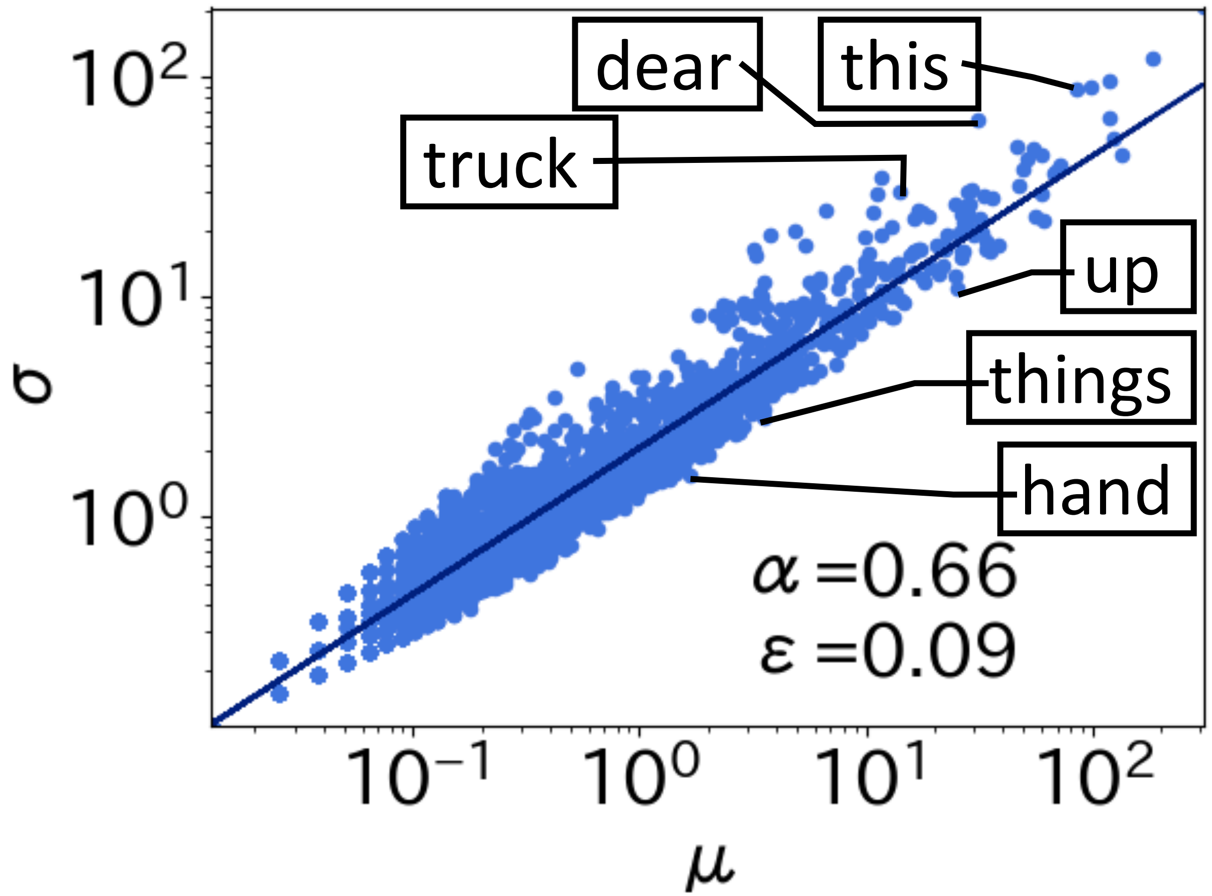
Thomas, English, CHILDES

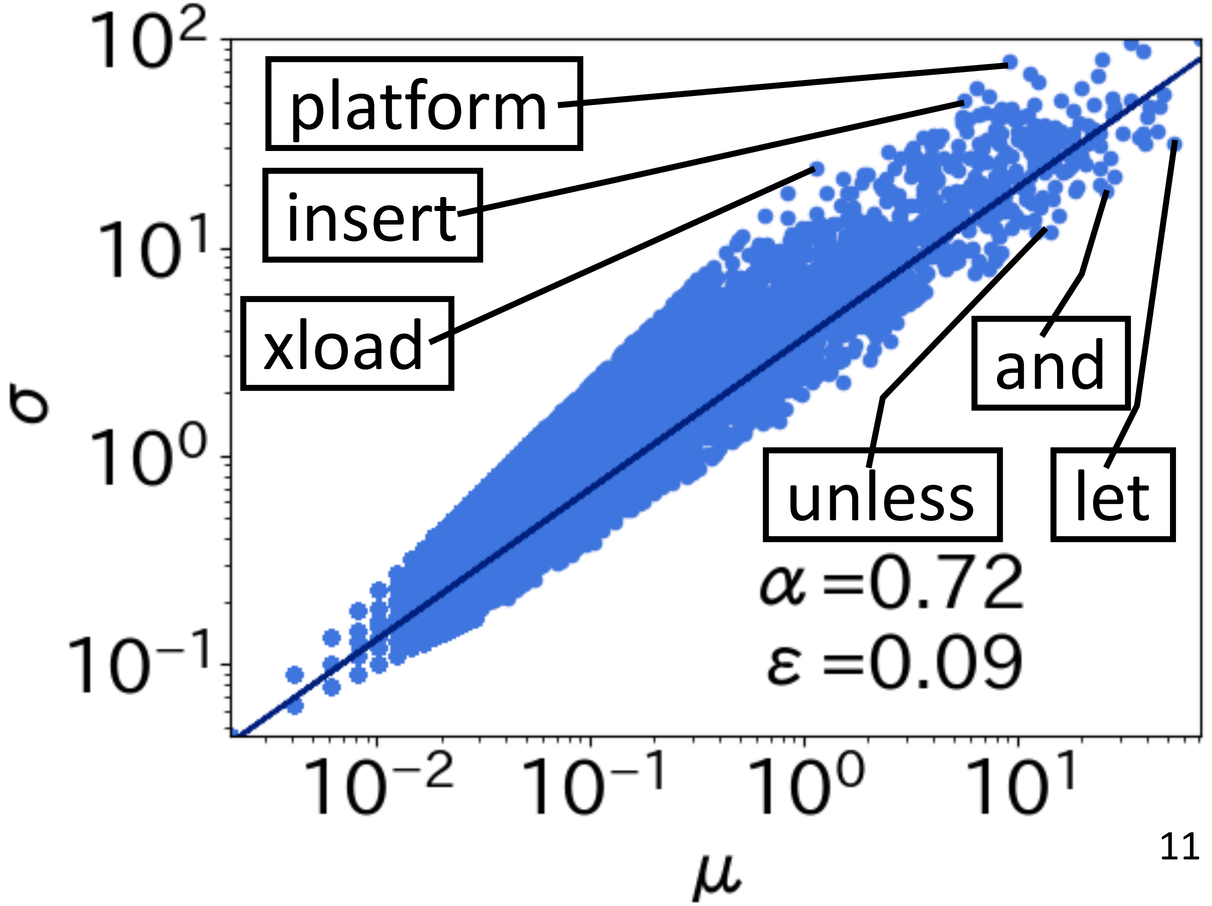
450k words (8.2k diff. words)

Programming source code

Lisp, crawled and parsed

3.7m words (160k diff. words)

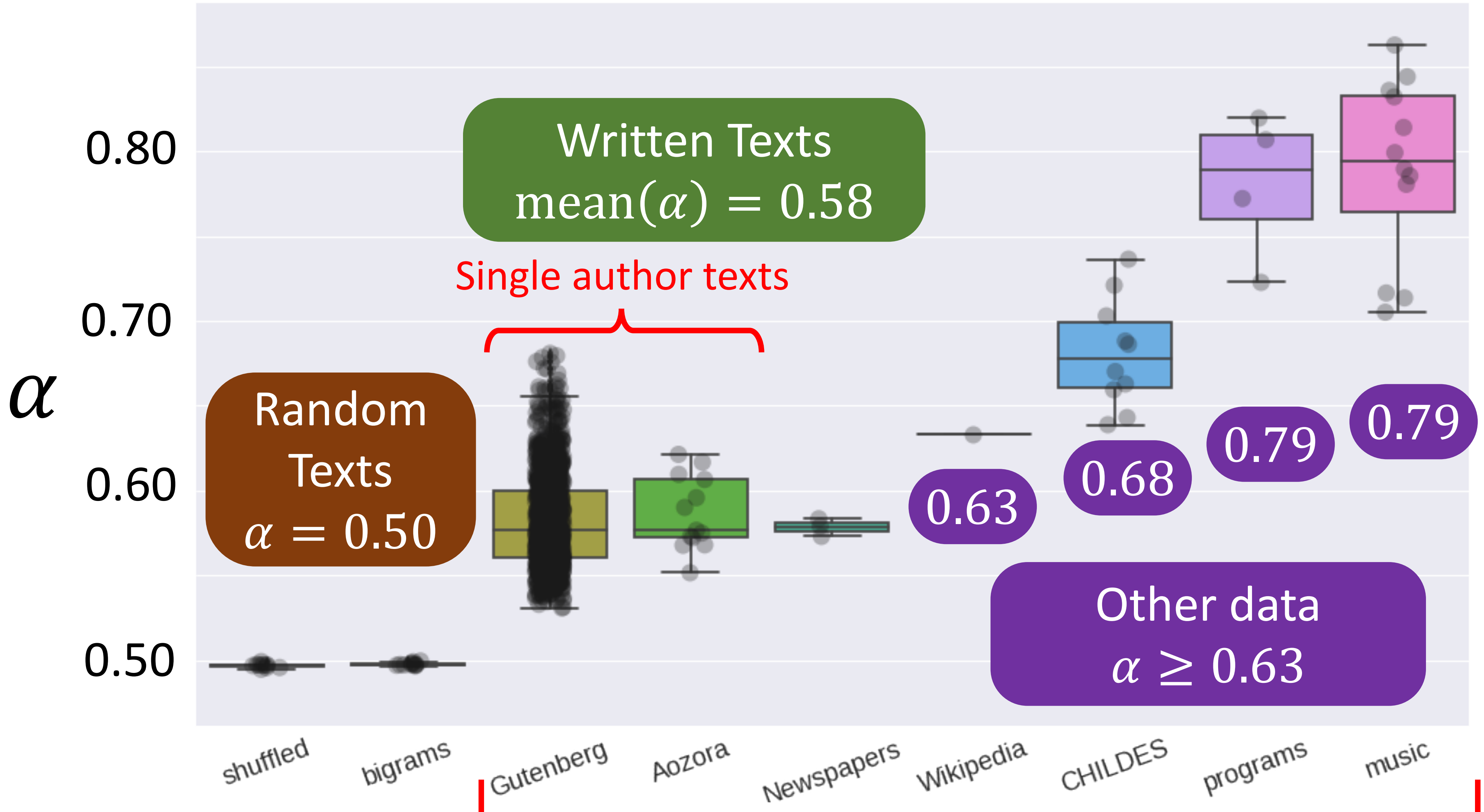




Datasets

Kind	Languages	Number of texts	Average size	Example
Gutenberg & Aozora (Long, single author)	14(En, Fr, …)	1142	311,483	‘Moby Dick’ ‘Les Miserables’
Newspapers	3 (En,Zh,Ja)	4	580,488,956	WSJ
Tagged Wiki	1 (En+tag)	1	14,637,848	enwiki8
CHILDES	10(En, Fr, …)	10	193,434	Thomas (English)
Music	-	12	135,993	Matthäus (Bach)
Program Codes	4	4	34,161,018	C++, Lisp, Haskell, Python

Taylor exponents of various data kind



None of the real texts showed the exponent 0.5

Summary thus far

- Taylor's law holds in vast fields including natural/social science
- Taylor's law also holds in languages and other linguistic related sequential data
- Taylor exponent shows the degree of co-occurrence among words
- Taylor exponent α differs among text categories

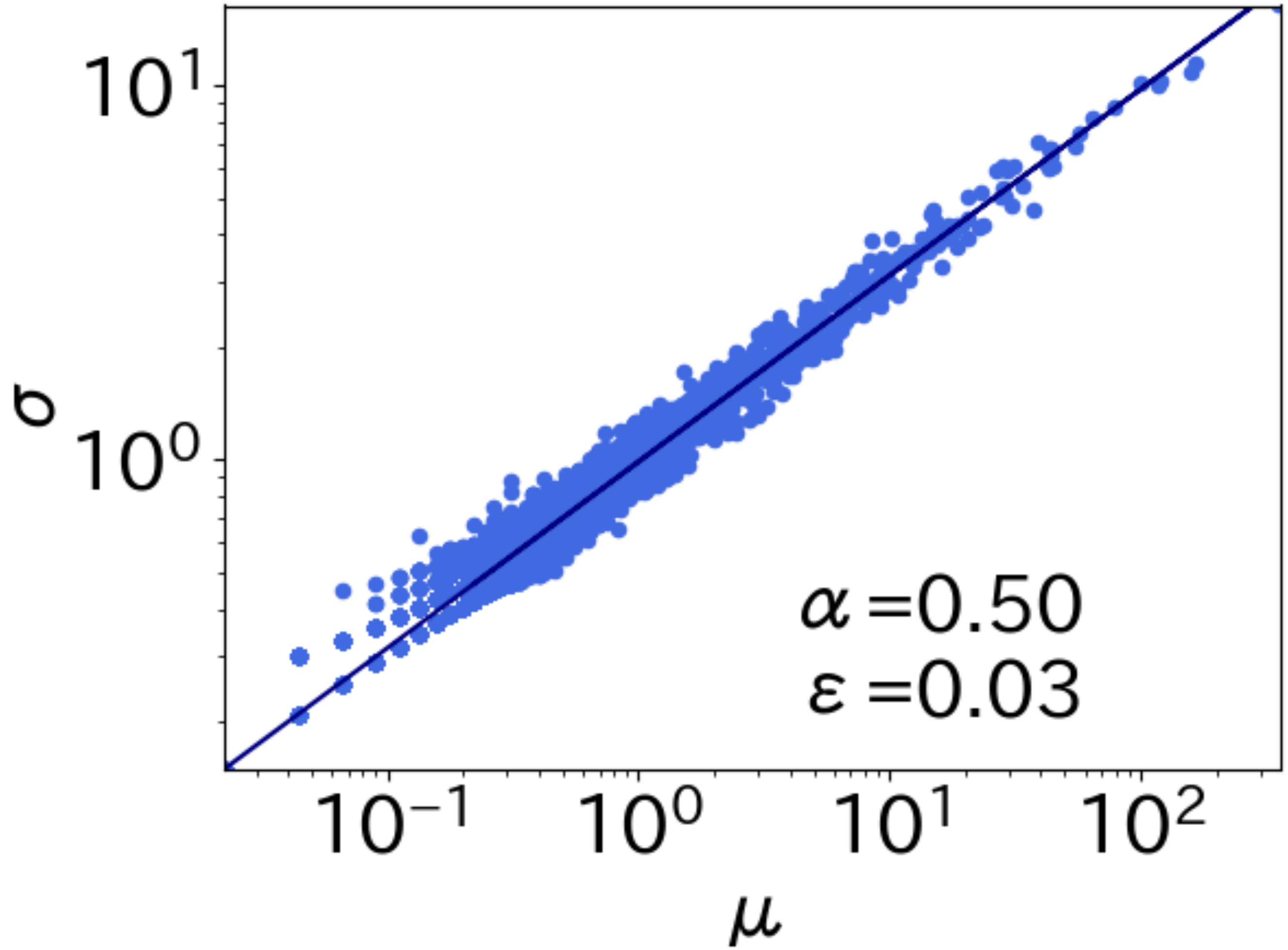
  (No such quality for Zipf's law, Heaps' law)

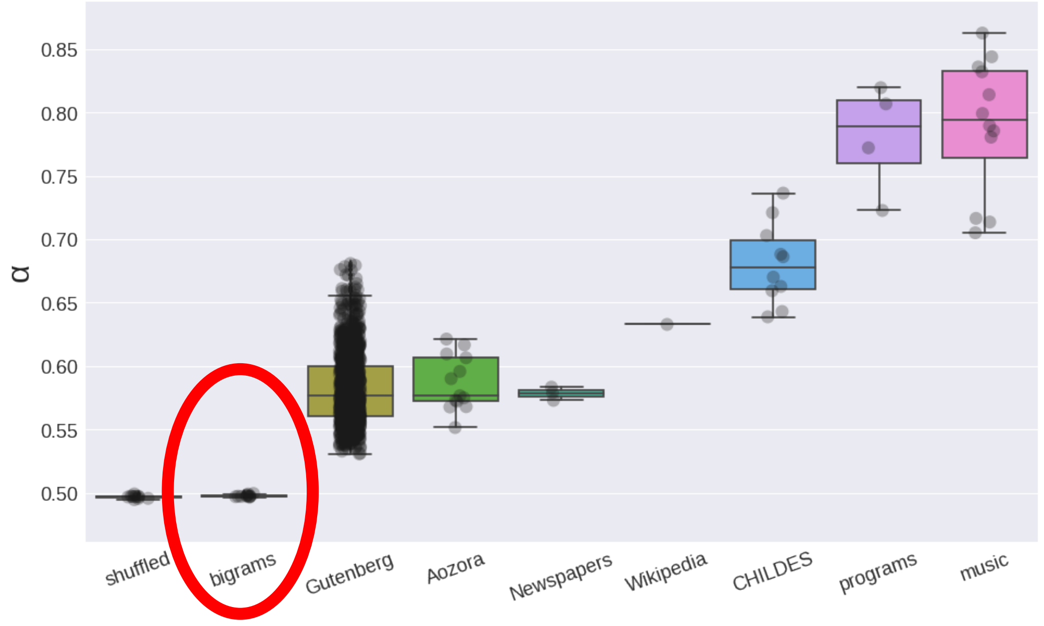
How can our results be useful?

$\Rightarrow$ Do machine generated texts produce $\alpha > 0.5$?

Machine generated text by *n*-grams

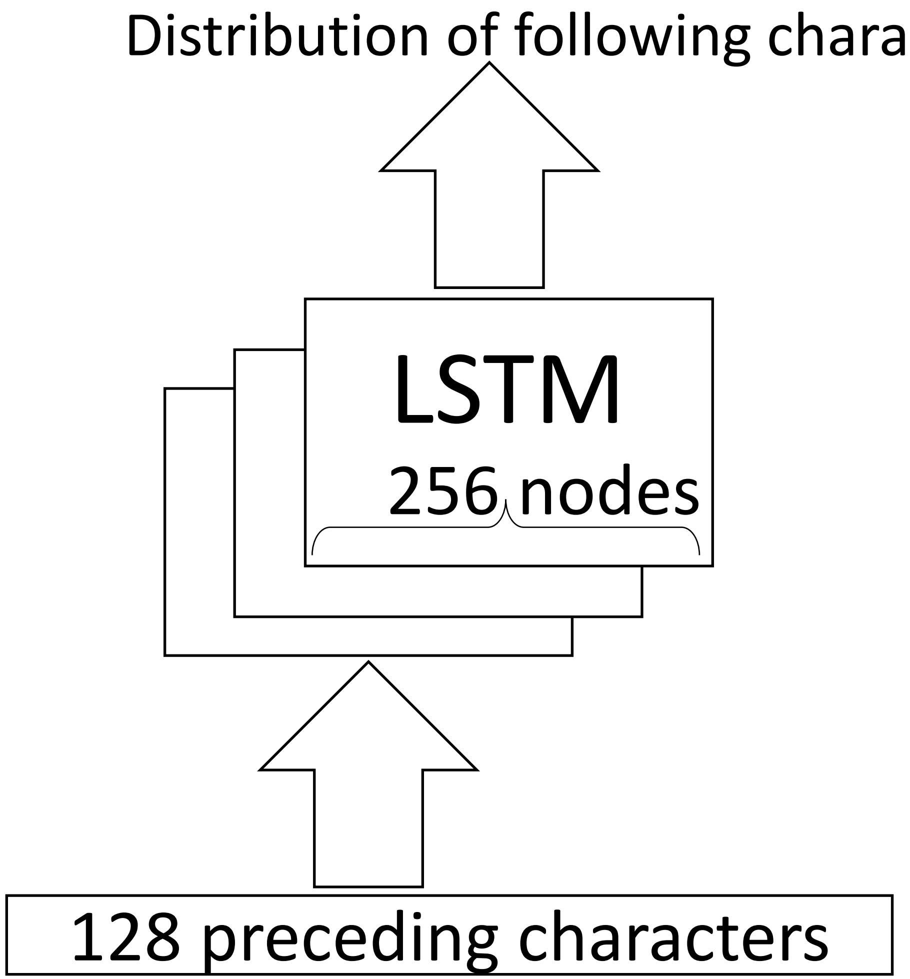
bigrams of Moby Dick





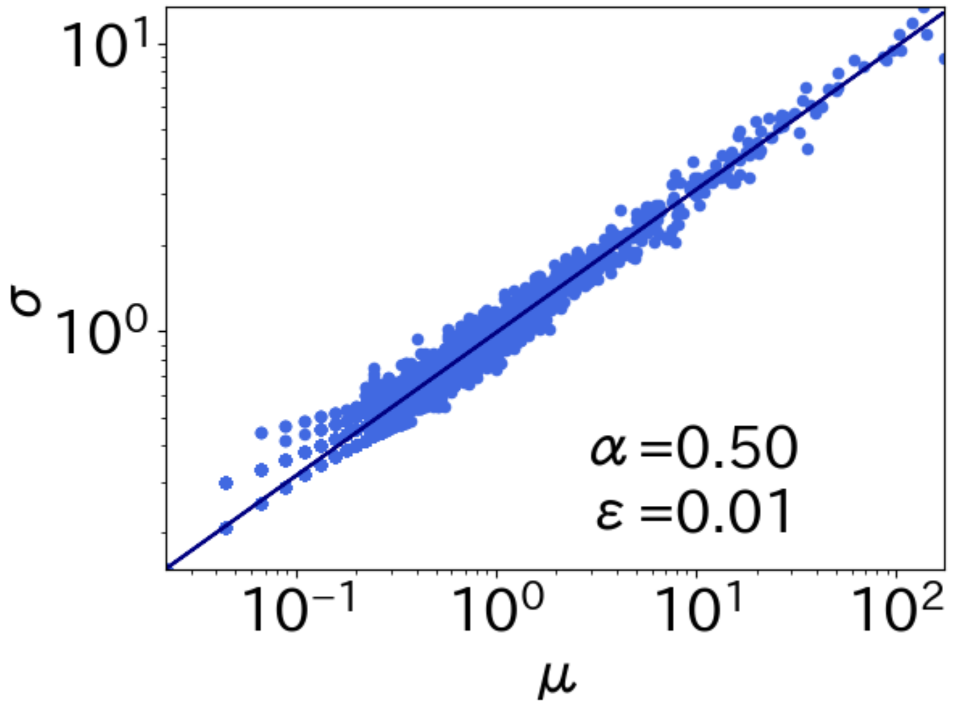
Machine generated texts by character-based LSTM language model

Stacked LSTM (3 LSTM layers)



Learning: Shakespeare by naive setting

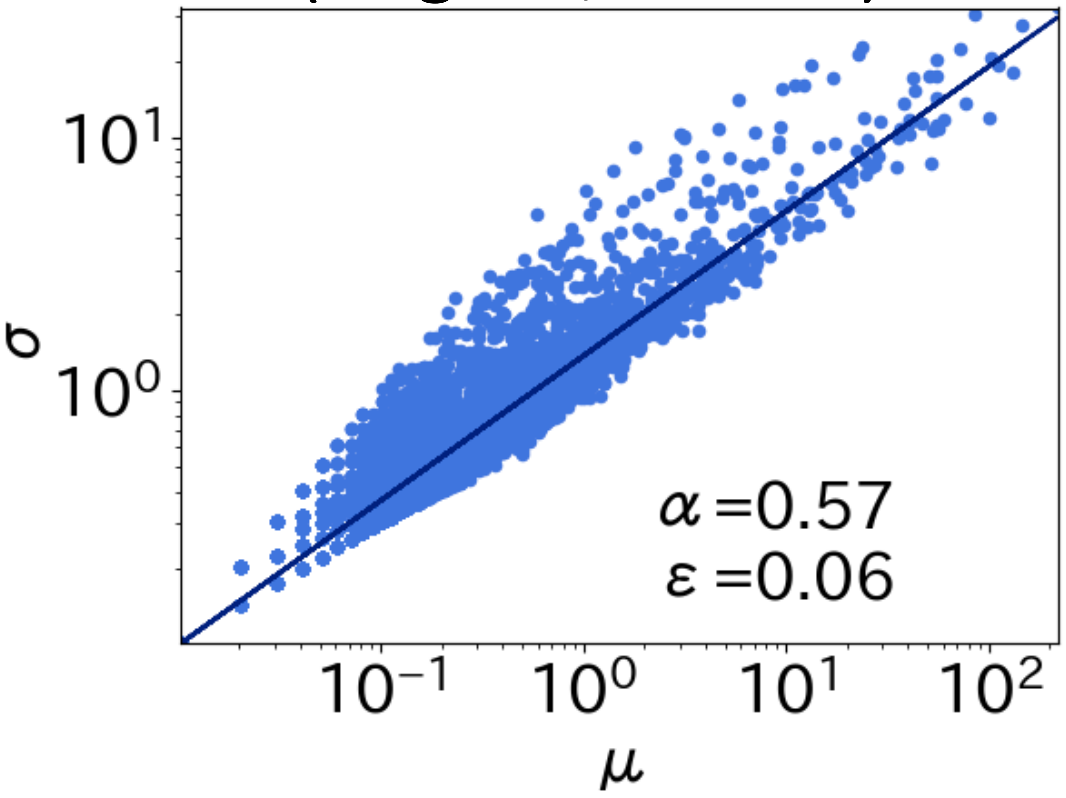
Generation: Probabilistic generation of succeeding characters (2 million characters)



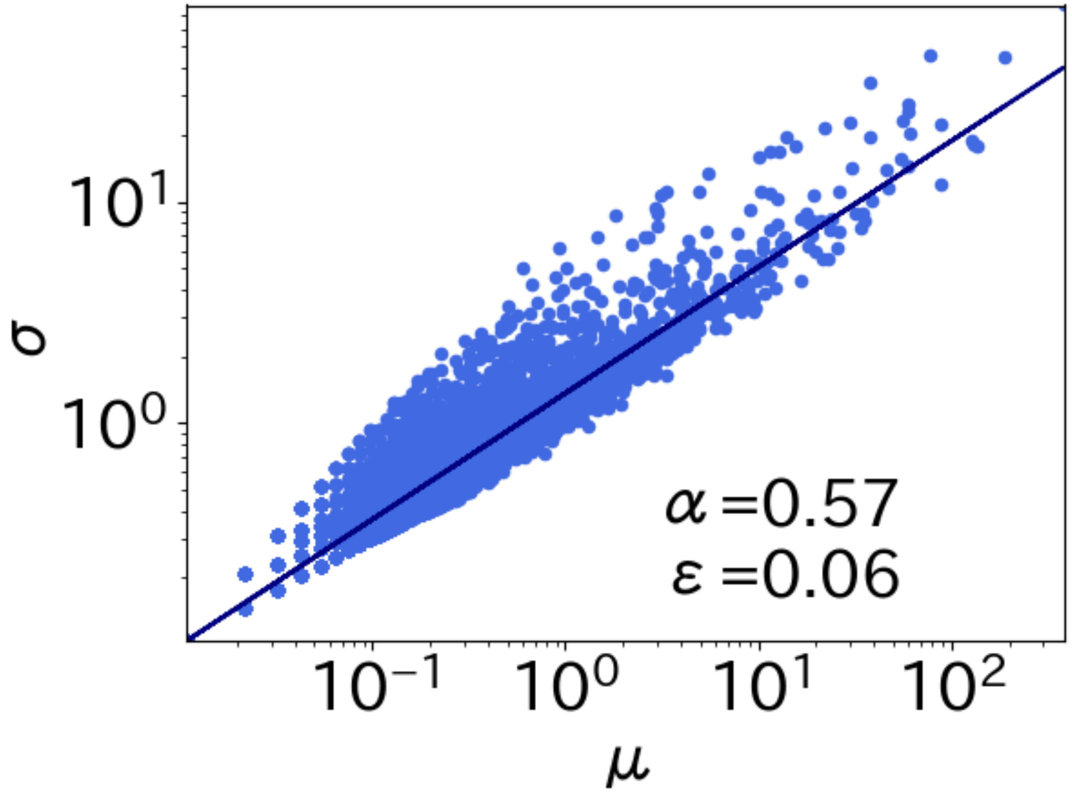
State-of the art models present different results (in another paper)

Texts generated by machine translation

Les Miserables (original, French)



Les Miserables translated by Google translator (in English)



Fluctuation that derives from the context is provided by the source text

Conclusion

- Taylor's law holds in vast fields including natural/social science
- Taylor's law also holds in languages and other linguistic related sequential data
- Taylor exponent shows the degree of co-occurrence among words
- Taylor exponent α differs among text categories

  (No such quality for Zipf's law, Heaps' law)

How can our results be useful?

$\Rightarrow$ Do machine generated texts produce $\alpha > 0.5$?

- The nature of $\alpha > 0.5$: context and long memory ← one limitation of CL
- Taylor analysis would possibly evaluate machine outputs
- Knowing mathematical characteristic of texts serve for language engineering

Thank you