Rare words and scaling laws in language

Eduardo Goldani Altmann

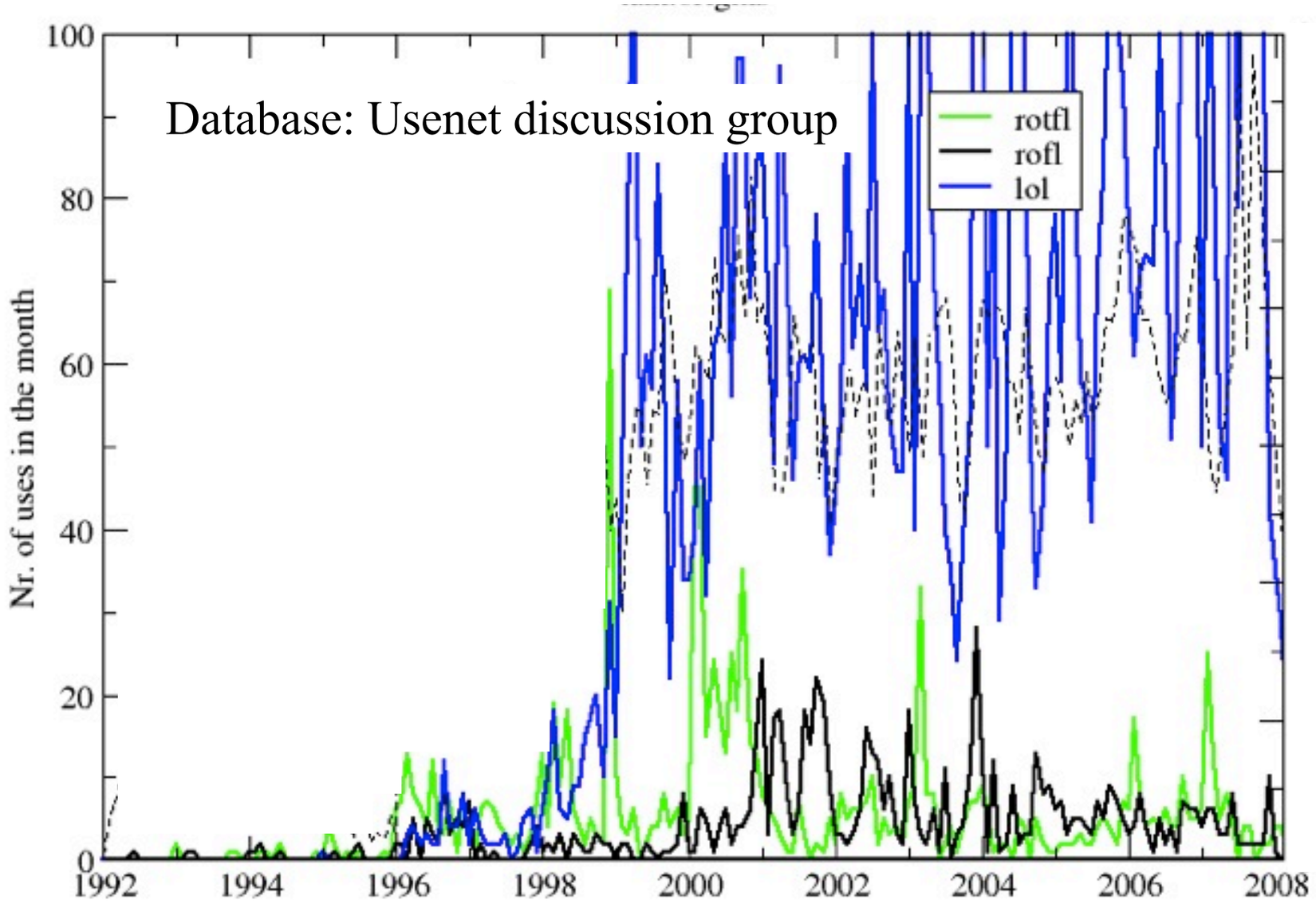
Max Planck Institute
for the Physics of Complex Systems,
Dresden, Germany

Paladin Conference, Rome
September 24, 2013

Statistical Analysis of Language

- Amount of produced data
  - Text messages: 10 B messages/day worldwide;
  - Twitter: 400 M tweets/day by 200M active users;
  - Wikipedia: 10 M contributors; 1 B words;
  - Google n-grams: 5 M books between [1520, 2000], 100 B words;

- Opportunity for applications (e.g., search engines, data mining) and scientific investigations (language as a lens on human activities and though).

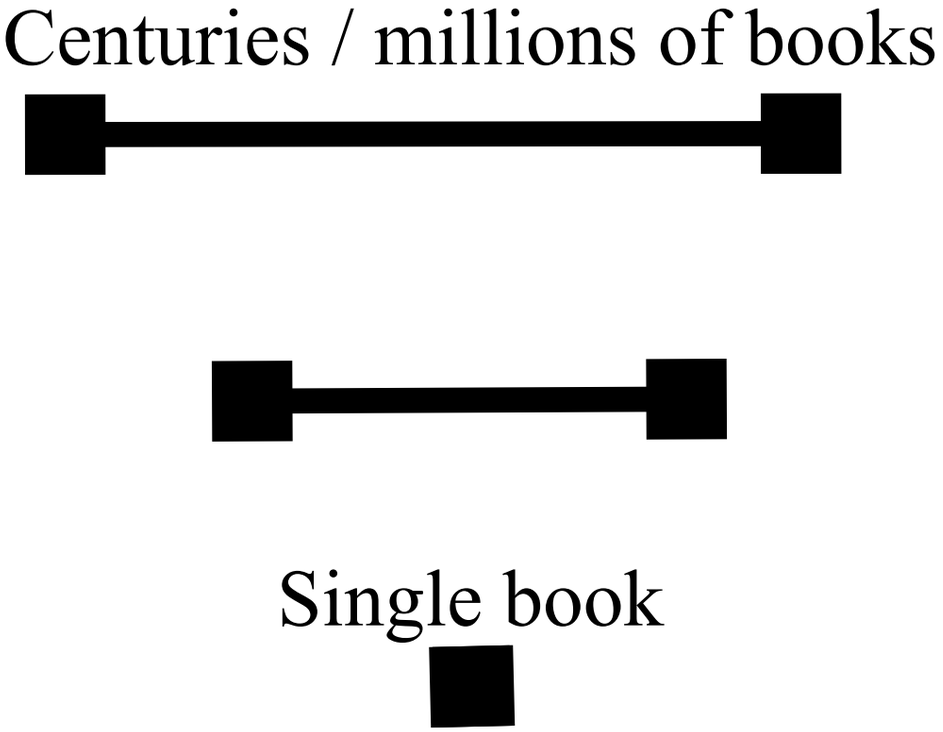


Plan:

1. Vocabulary Growth

2. Innovations and Change

3. Text Analysis

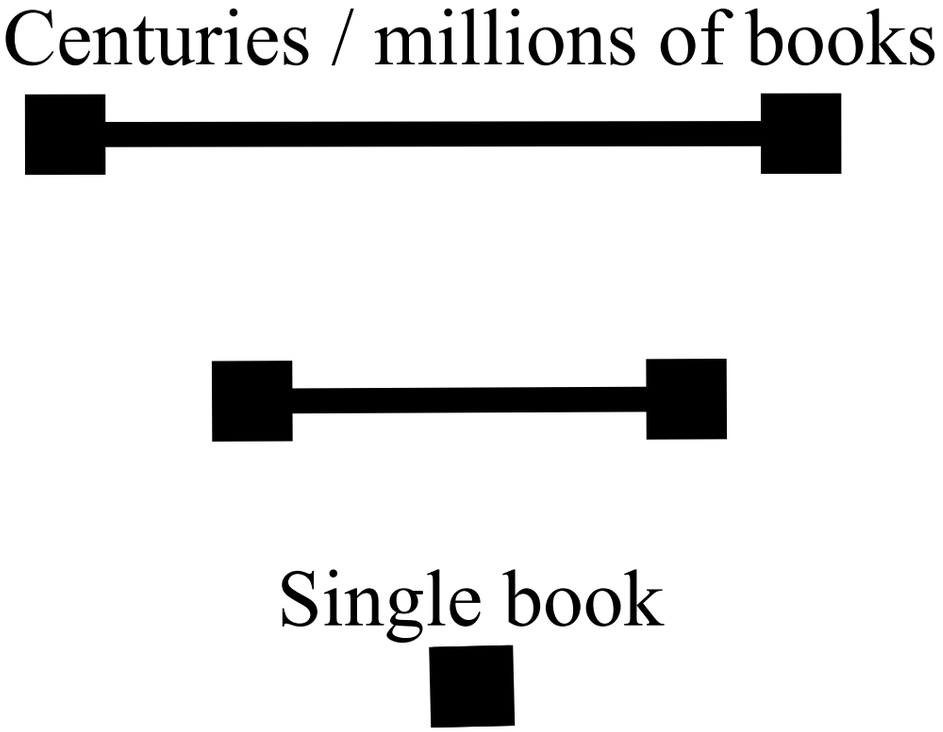


Plan:

1. Vocabulary Growth

2. Innovations and Change

3. Text Analysis



Martin Gerlach and E. G. Altmann,
Stochastic model for the vocabulary growth in natural languages,
Phys. Rev. X 3, 021006 (2013)

Motivation

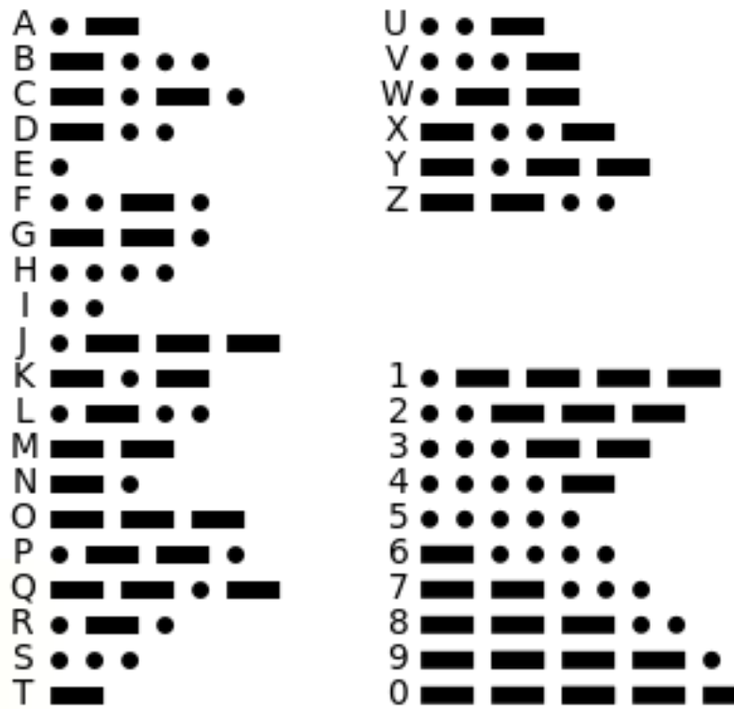
Samuel Morse (1791-1872)
[Wikipedia]

message

the ... also .. prince ...

e t a i .. q y ..

rank=1 2 3 ...



International Morse Code
[Wikipedia]

"The dictionary or vocabulary consists of words alphabetically arranged and regularly numbered... so that each word in the language has its telegraphic number..."

[Morse's first telegraph patent as cited by J. Gleick, *The Information*]

How the vocabulary grows in time and with database size?

Motivation: invert indexing

Vocabulary size
=
memory allocation

	page 1	page 2	page 3	page 4	page 5	page 6
“the”						
“it”						
…						
word n						
…						
word N						

Motivation: vocabulary of a language?

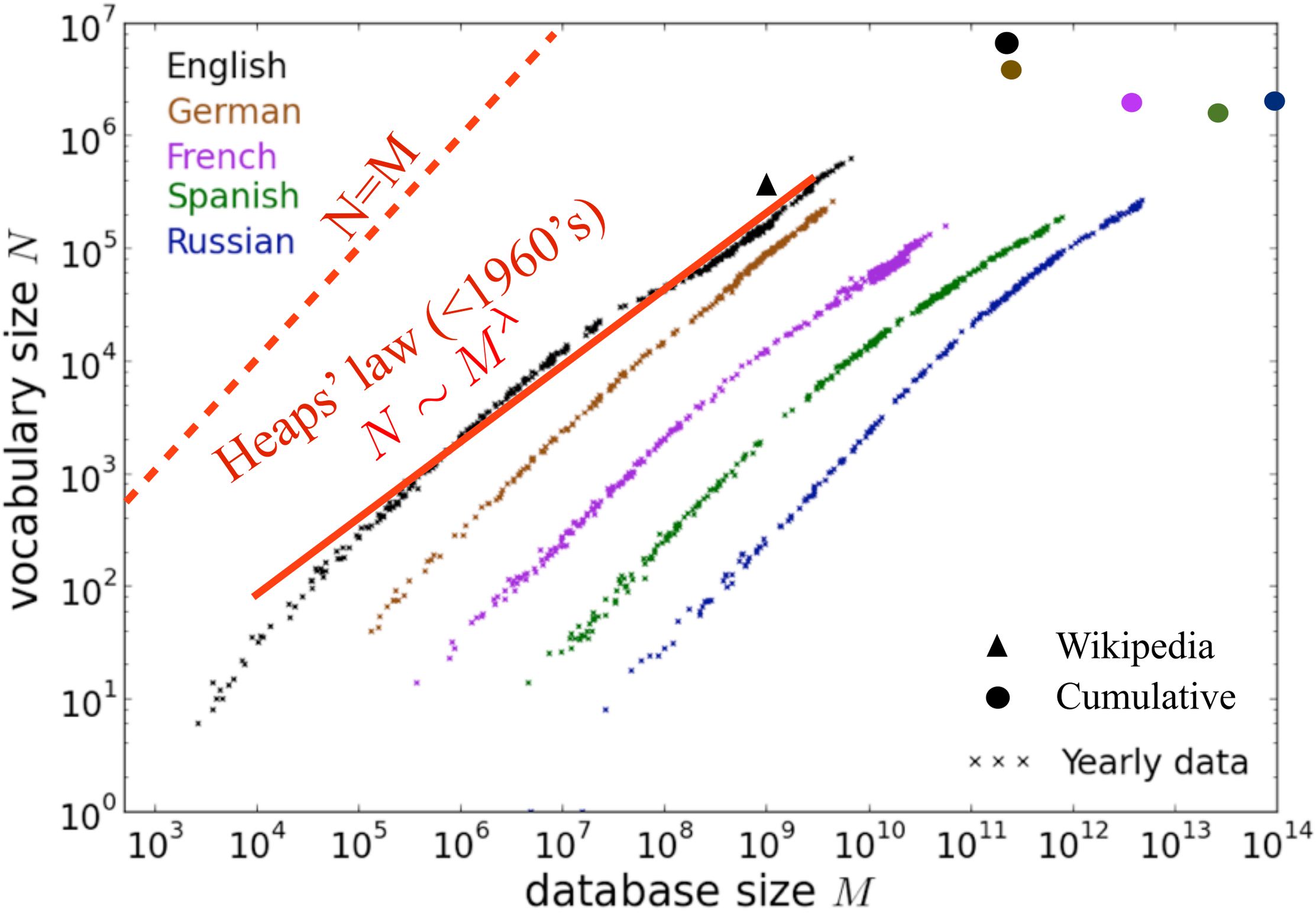
Report on the state of the German language (March 2013) German Academy for Language and Literature Union of German Academies of Sciences and Humanities			
Year	1905-1914	1948-1957	1995-2004
# distinct words	3,715,000	5,045,000	5,238,000

Quantitative Analysis of Culture Using Millions of Digitized Books Michel et. al., Science (2011) [*English*]			
Year	1900	1950	2000
# distinct words	544,000	597,000	1,022,000

Problem: role of database size?

Vocabulary growth with database size

Limit vocabulary?

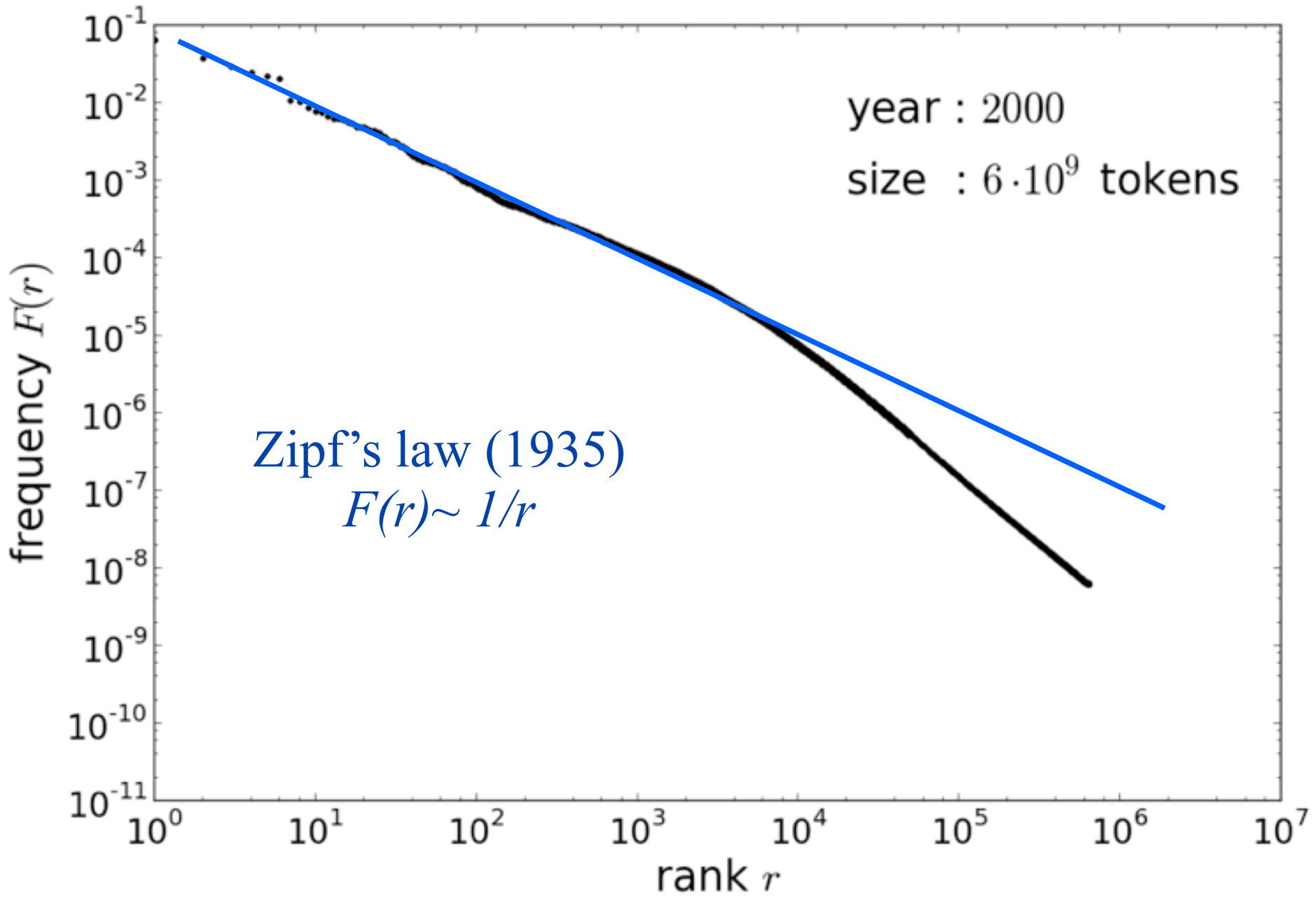


Descriptive model

Simple mode: usage of each word follows a Poisson process with fixed frequency

$$\langle N(M) \rangle = \sum_r 1 - e^{-F(r)M}$$

where *F(r)* is the frequency of the *r*-th most frequent word (*r* = rank).

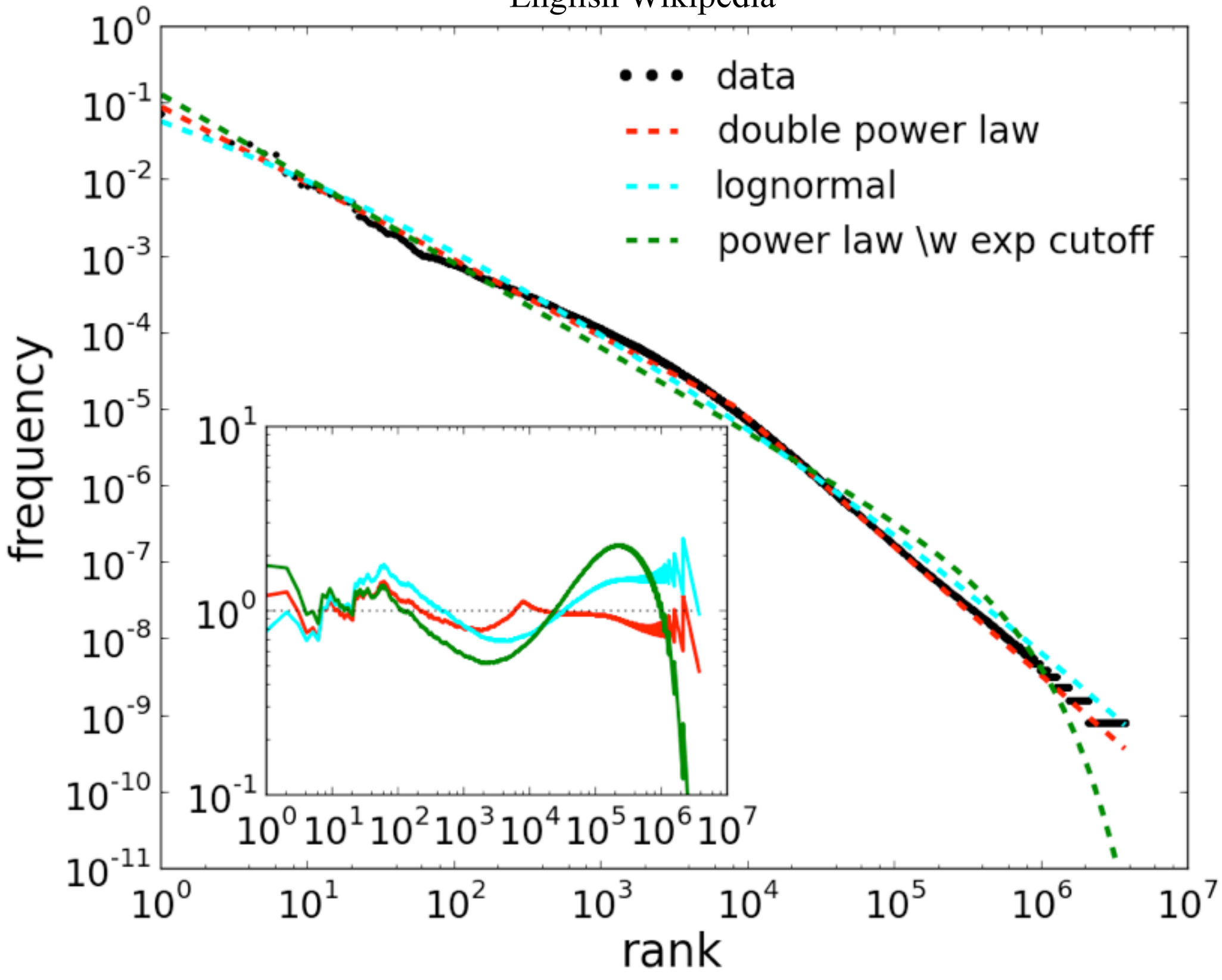


Generalized Zipf's law

i	distribution	$F(r;\Omega)$	set of parameters Ω
1	Power-Law	$Cr^{-\gamma}$	γ
2	Shifted Power-Law	$C(r+b)^{-\gamma}$	γ, b
3	Power-Law with Exponential cutoff (beginning)	$C\exp(-b/r)\,r^{-\gamma}$	γ, b
4	Power-Law with Exponential cutoff (tail)	$C\exp(-br)\,r^{-\gamma}$	γ, b
5	Log-normal	$Cr^{-1}\exp\left(-\frac{1}{2}\left(\ln r-\mu\right)^2/\sigma^2\right)$	μ, σ
6	Weibull	$Cr^{\gamma-1}\exp\left(-br^{-\gamma}\right)$	γ, b
7	Double Power-Law	$C\begin{cases} r^{-1}, & r\le b \\ b^{\gamma-1}r^{-\gamma} & r>b, \end{cases}$	γ, b

Generalized Zipf's law

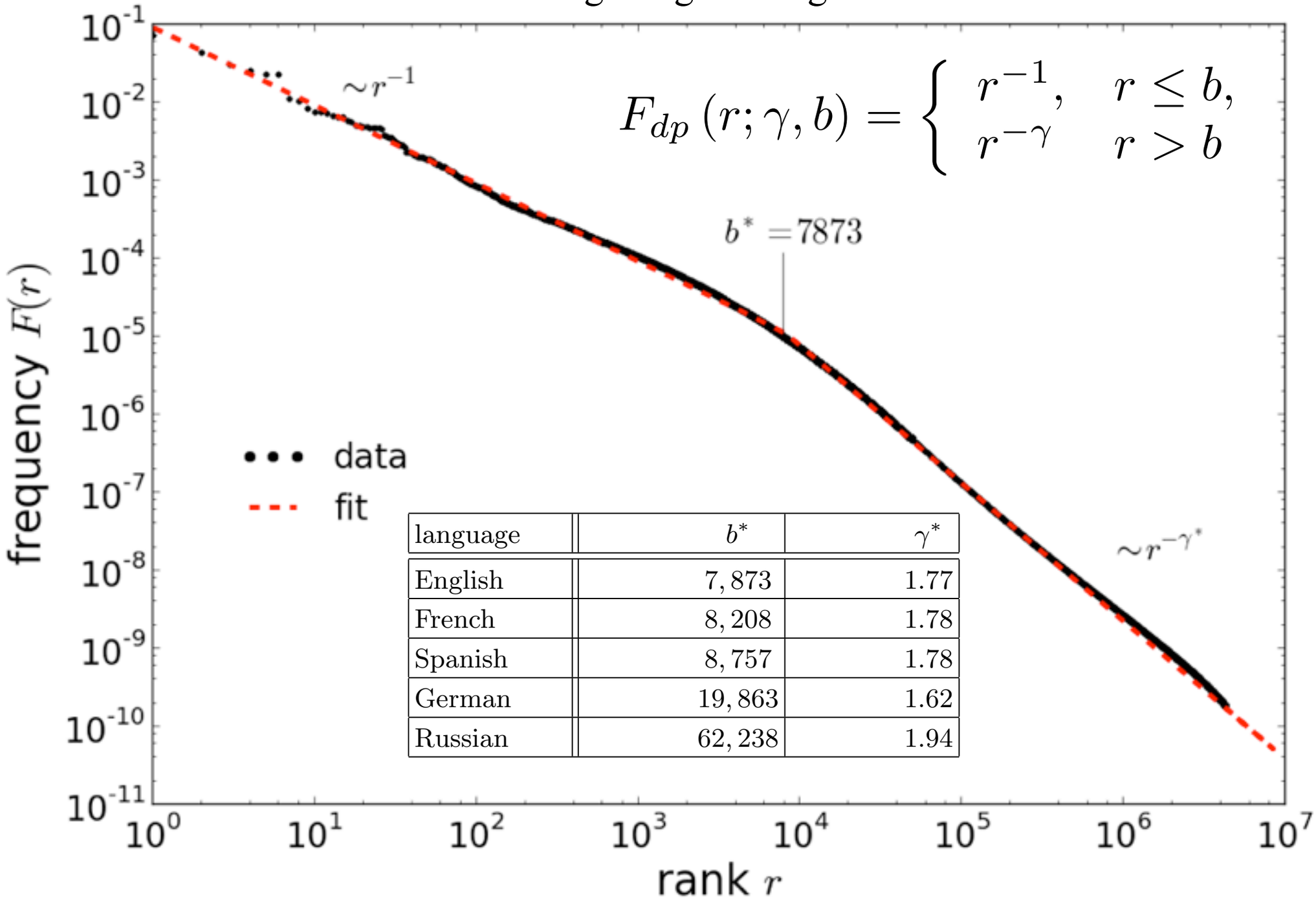
English Wikipedia



Generalized Zipf's law

Google n-gram English

$$F_{dp}(r;\gamma,b) = \begin{cases} r^{-1}, & r \le b, \\ r^{-\gamma} & r > b \end{cases}$$



language	b^*	γ^*
English	7,873	1.77
French	8,208	1.78
Spanish	8,757	1.78
German	19,863	1.62
Russian	62,238	1.94

Descriptive model

Simple mode: usage of each word follows a Poisson process with fixed frequency

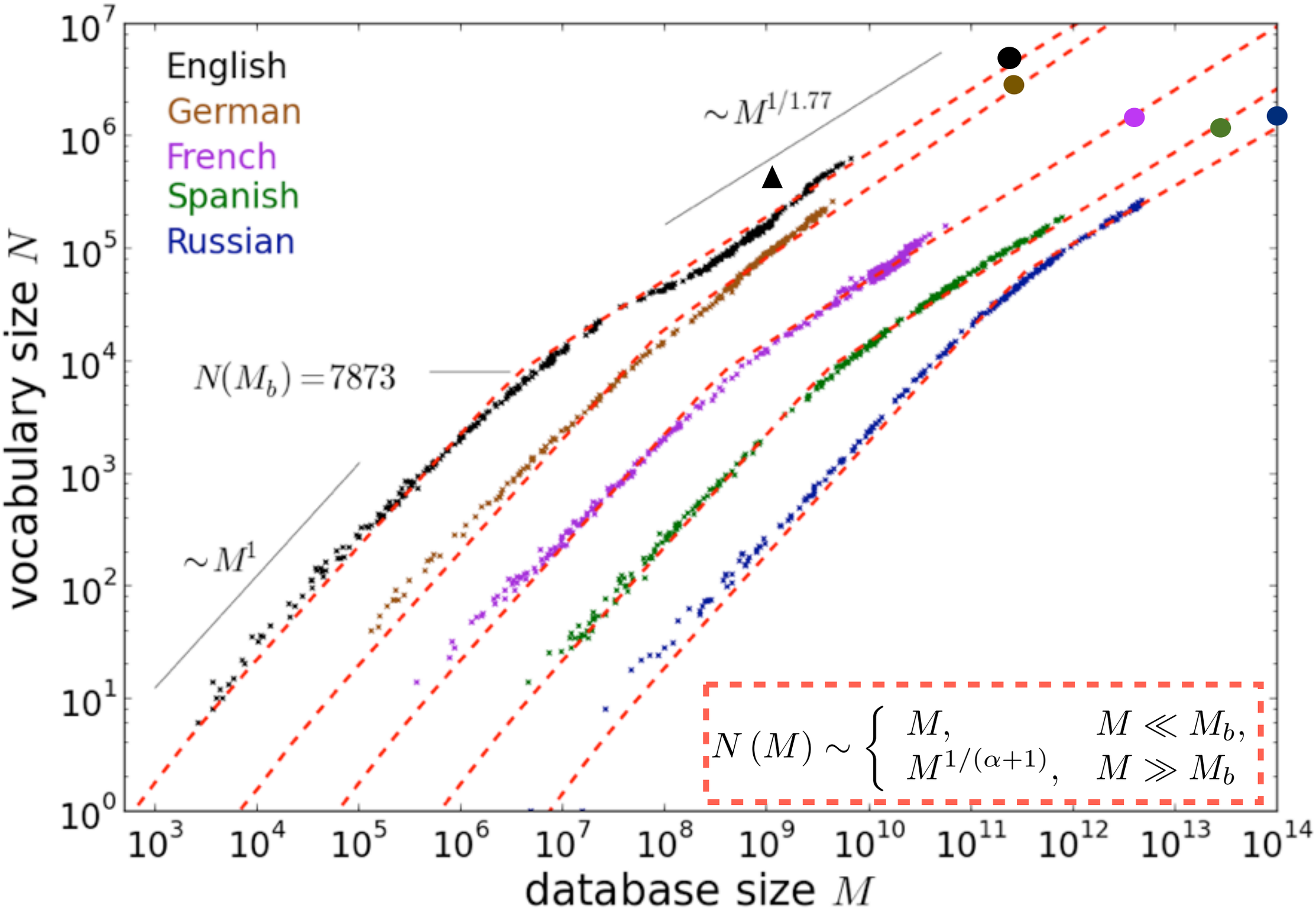
$$\langle N(M)\rangle = \sum_r 1 - e^{-F(r)M}$$

where F(r) is the frequency of the r-th most frequent word (r = rank).

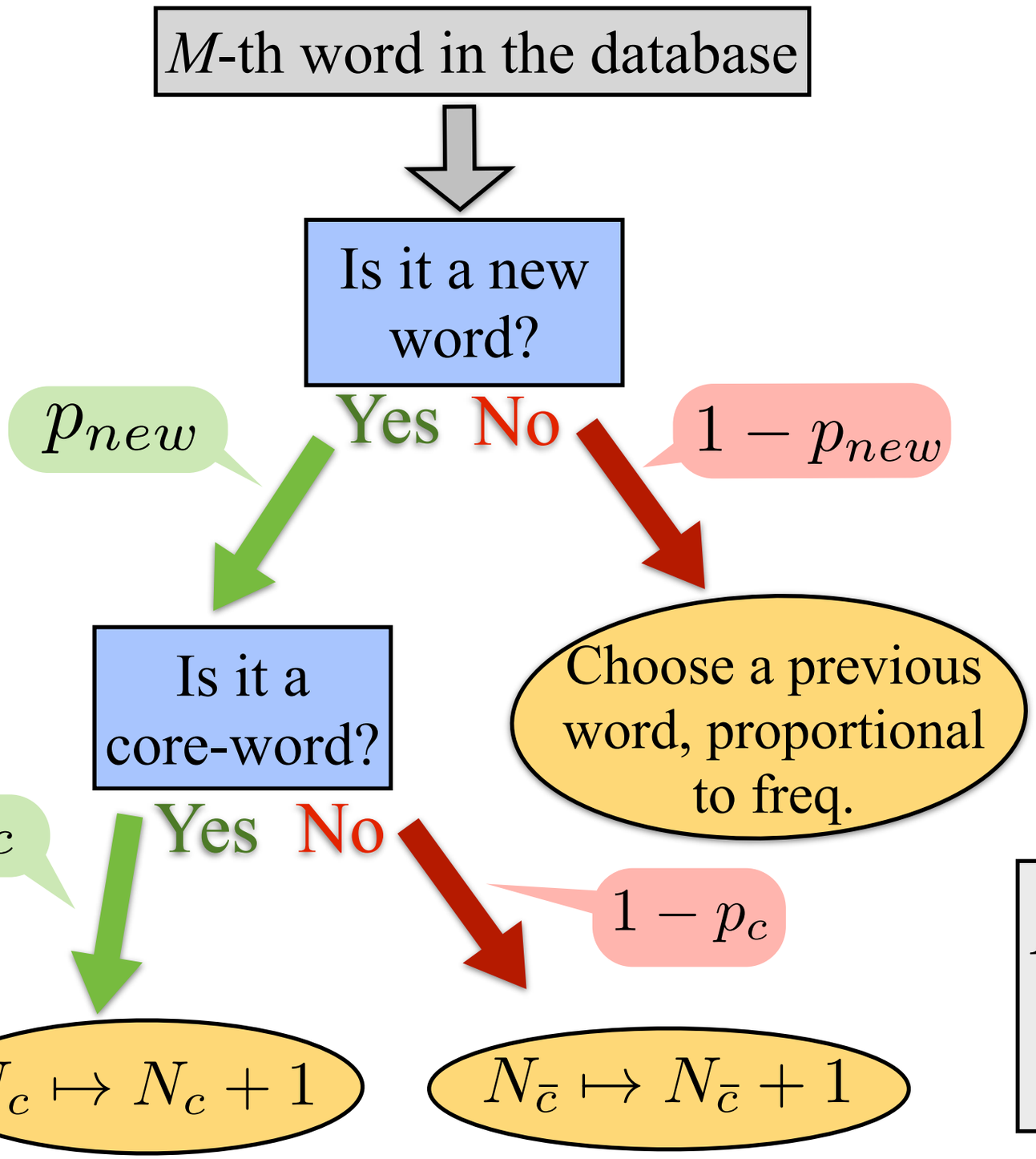
$$F_{dp}(r;\gamma,b) = \begin{cases} r^{-1}, & r \le b, \\ r^{-\gamma} & r > b \end{cases}$$

$$N_{dp}(N_c) = \begin{cases} M, & M \ll M_b, \\ M^{1/\gamma}, & M \gg M_b \end{cases}$$

Descriptive model



Generative model (Yule-Simon type)



Assumptions:

1. Core vocabulary is finite:

$$N_c \leq N_c^{max} \Rightarrow p_c \to 0$$

2. p_{new} decays with N:

$$p_{new} \mapsto p_{new}\,(1 - \alpha/N)$$

$$N\,(M) \sim \begin{cases} M, & M \ll M_b, \\ M^{1/(\alpha+1)}, & M \gg M_b \end{cases}$$

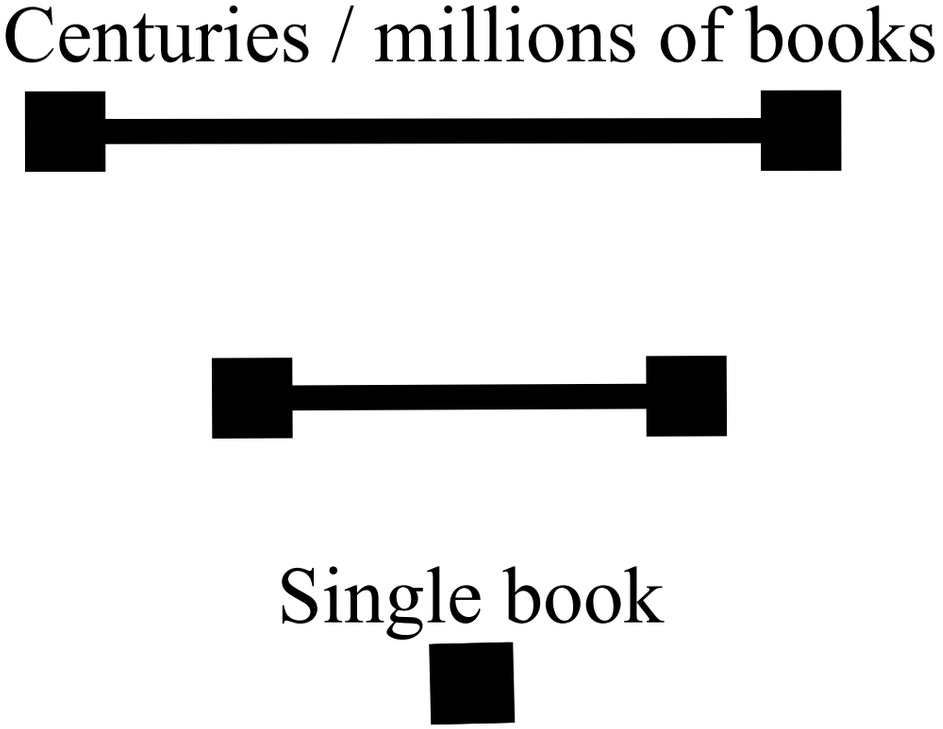
with $N(M_b) = N_c^{max}$

Plan:

1. Vocabulary Growth

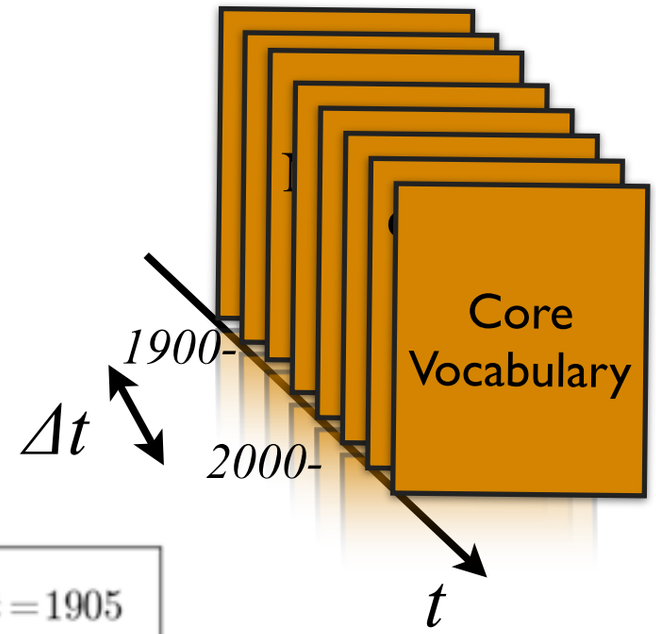
2. Innovations and Change

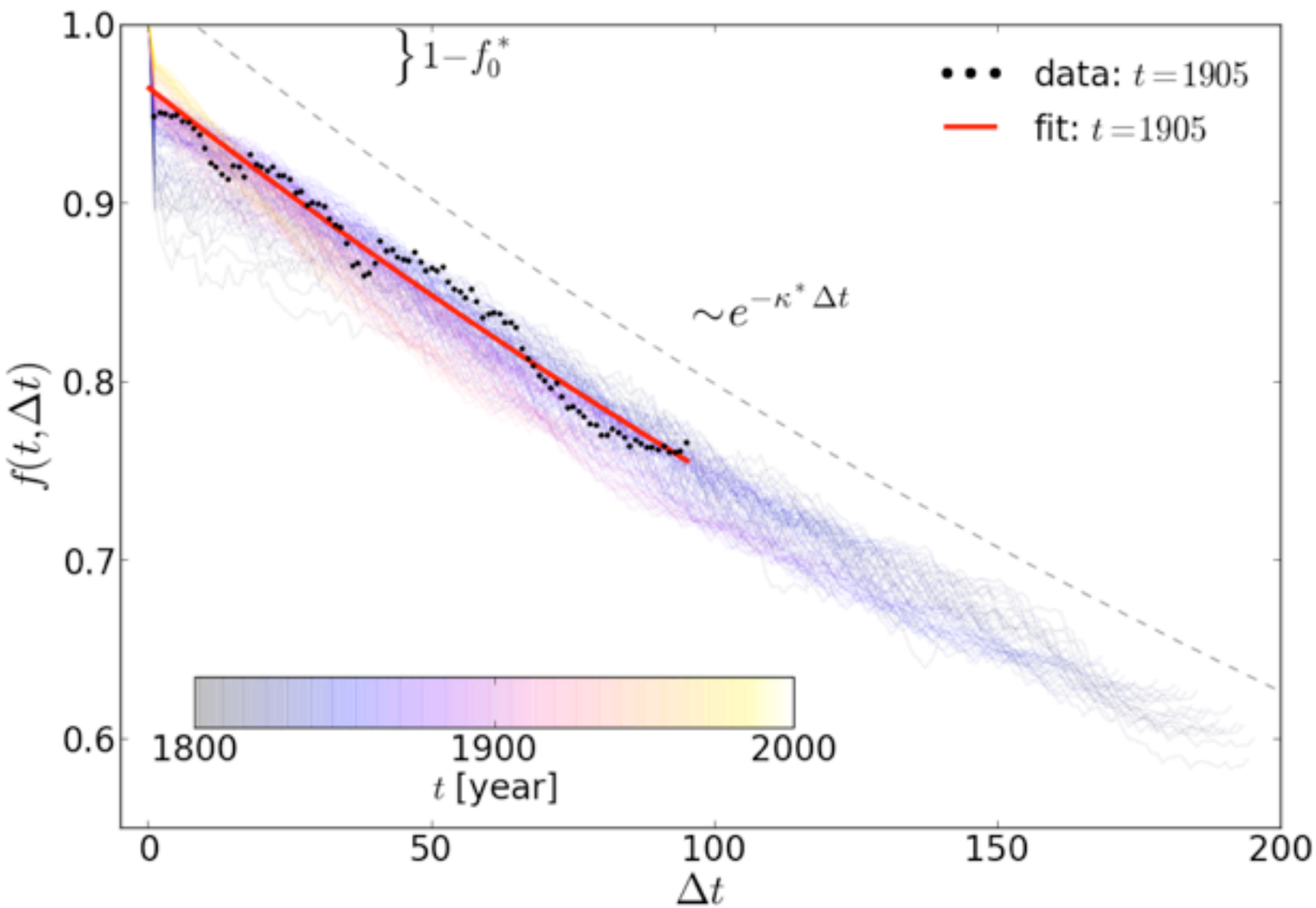
3. Text Analysis



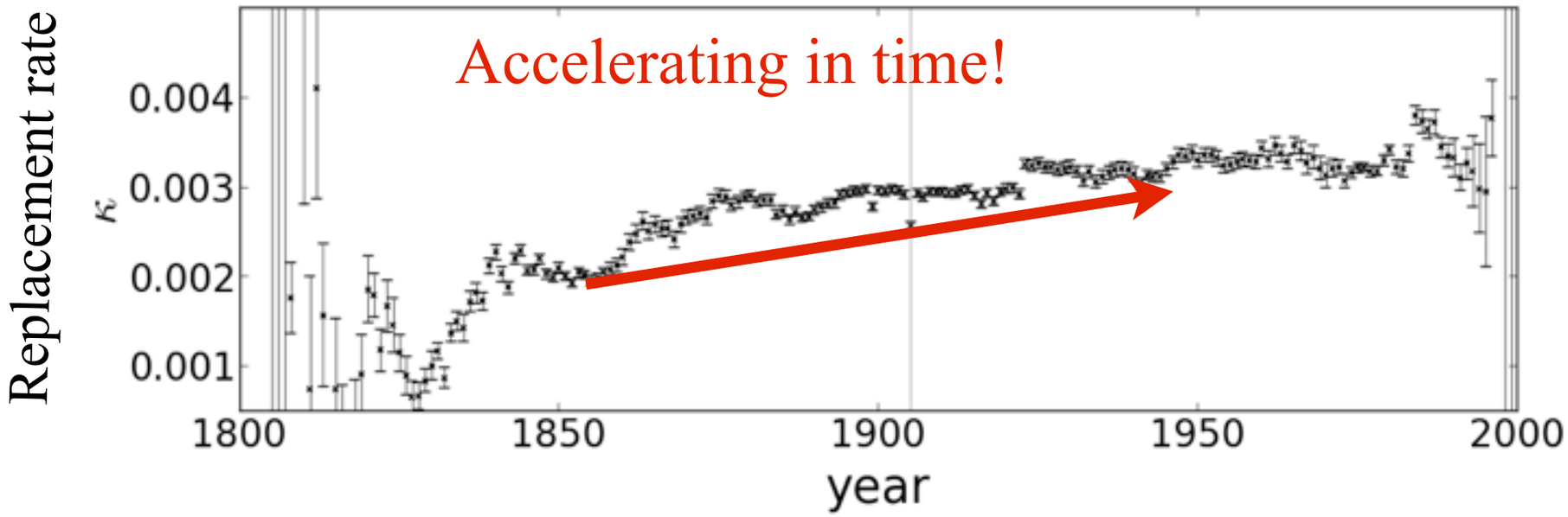
Change in the core vocabulary

f(t,Δt): fraction of core words at time *t* which remain core at time *t*+*Δt*





Change in the core vocabulary



Change in the core vocabulary

1900

majesty, doubtless, furnished, monsieur, Napoleon, hitherto

Most frequent replaced words

2000

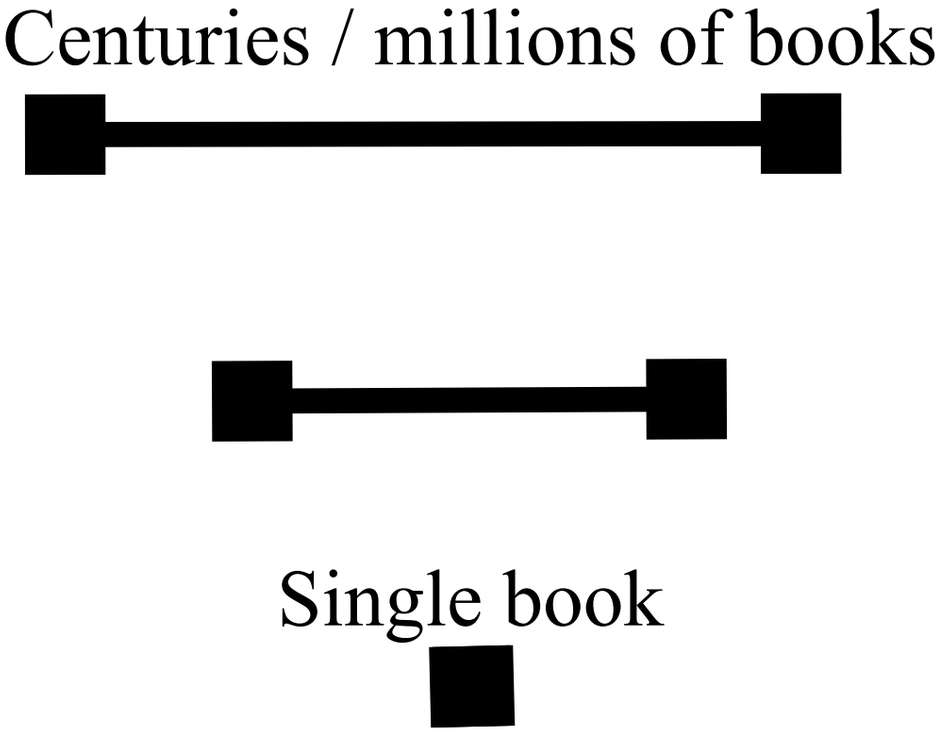
cultural,context, technology, programs, environmental, computer

Plan:

1. Vocabulary Growth

2. Innovations and Change

3. Text Analysis



E. G. Altmann, Giampaolo Cristadoro, Mirko Degli Esposti, "*On the origin of long-range correlations in texts*", PNAS 109, 11582 (2012)

Observation

Symbolic sequence *s*

War and Peace, by Leo Tolstoy
BOOK ONE: 1805
CHAPTER I
"Well, Prince, so Genoa and Lucca are now just family estates of the Buonapartes. But I warn you, if you don't tell me that this means war, if you still try to defend the infamies and horrors perpetrated by that Antichrist--I really believe he is Antichrist--I will have nothing more to do with you and you are no longer my friend, no longer my 'faithful slave,' as you call yourself! But how do you do? I see I have frightened you--sit down and tell me all the news."

. . .

$f(s)$

Numeric sequence

01001000011010000011010100010
0110010100000
001001000
0100000100100100101101000010010101010
0100001010001001010010000100110101001000100 1
0100001100000110010000100001000100001 0
0110000100100011000100000000100101000 01
1001011001000010010000100100101000000 10
1001000100000011000010110100101001001000100001
0100001010001001000010100100100100001101 00
0110101001001001000000011000010010010000
0110001000101010001101000011001000010001 0
0100110010101100001010001000101000110010
0100010000100001010000100 1000

. . .

text s=War and Peace by Leo Tolstoy Book one 1805 Chapter 1 Well Prince

vowels=0100100011010000011001000100110010000000000100100000100000100 1...

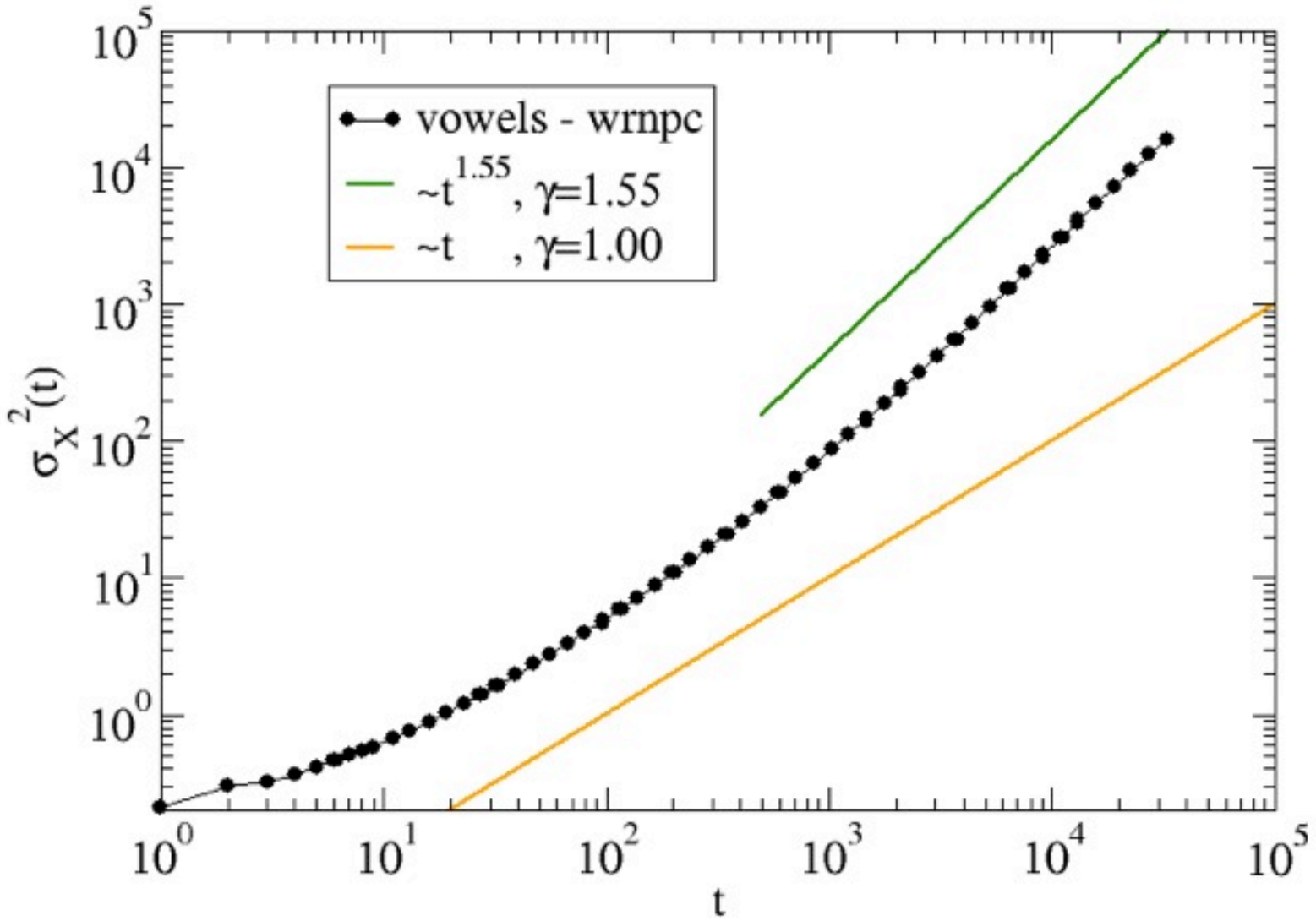
t

Random walk X(t):

0 => left
1 => right

$$\sigma_X^2(t) := \langle X(t)^2 \rangle - \langle X(t) \rangle^2$$

Transport and long correlations

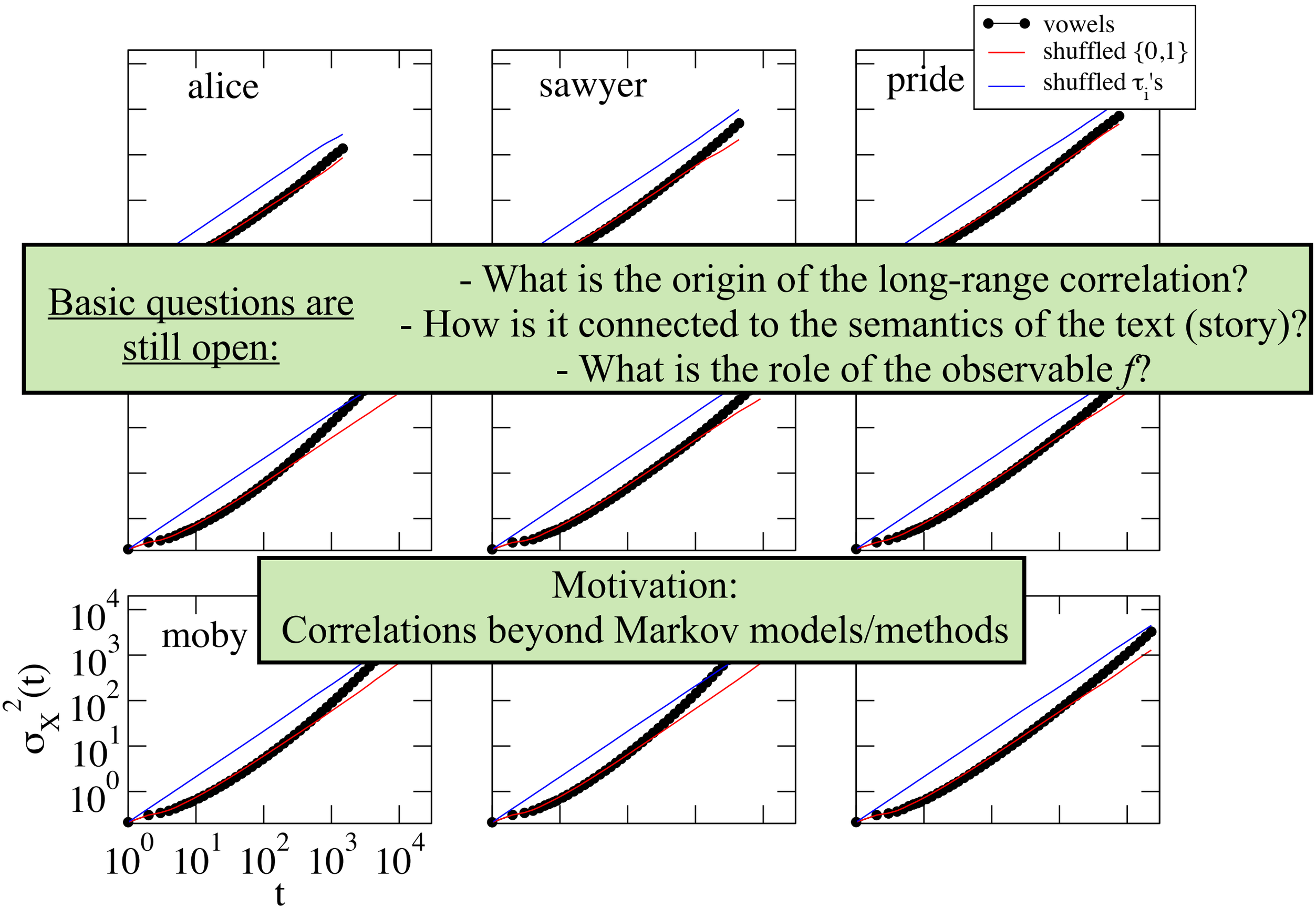


Super-diffusion: $\sigma_X^2(t) \simeq t^\gamma, \quad 1 < \gamma < 2$

Long-range correlation: $C_{\mathbf{x}}(t) \simeq t^{-\beta}, \quad 0 < \beta < 1$

$$\gamma = 2 - \beta$$

Different books



Basic questions are still open:

- What is the origin of the long-range correlation?
- How is it connected to the semantics of the text (story)?
- What is the role of the observable *f*?

Motivation:
Correlations beyond Markov models/methods

Burstiness and correlation

Inter-event times: **x**=1001000111 01...

$$\tau = 4$$

$$\tau_k = 3, 4, 1, 2, \ldots$$

Transport

$$\sigma_X^2(t) := \langle X(t)^2 \rangle - \langle X(t) \rangle^2 \simeq t^\gamma,$$

Power spectrum

$$S(\omega) := \int_{-\infty}^{+\infty} C_{\mathbf{x}}(t) e^{2\pi i t \omega} dt$$

short range $\quad \gamma = 1$

long range $\quad 1 < \gamma \leq 2$

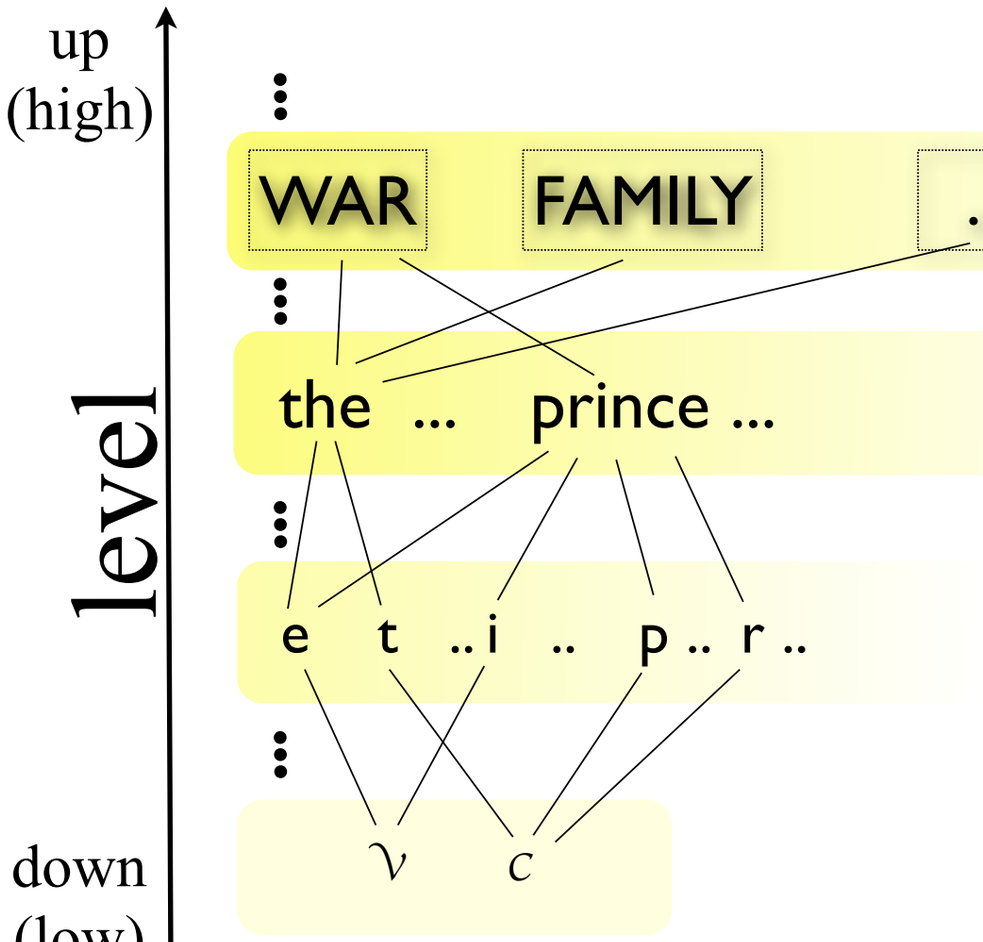
$$S(0) = \frac{\sigma_\tau^2}{\langle \tau \rangle^3} \left(1 + 2 \sum_k C_\tau(k) \right)$$

$$S(0) \to \infty$$

Burstiness $\quad \dfrac{\sigma_\tau}{\langle \tau \rangle} \to \infty$

Correlation in τ_k $\quad \displaystyle\sum_{k=1}^{\infty} C_\tau(k) \to \infty$

Model

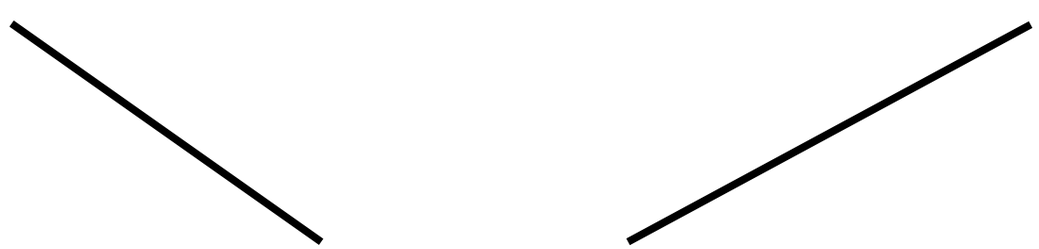


e=0100100011010000011001 00...

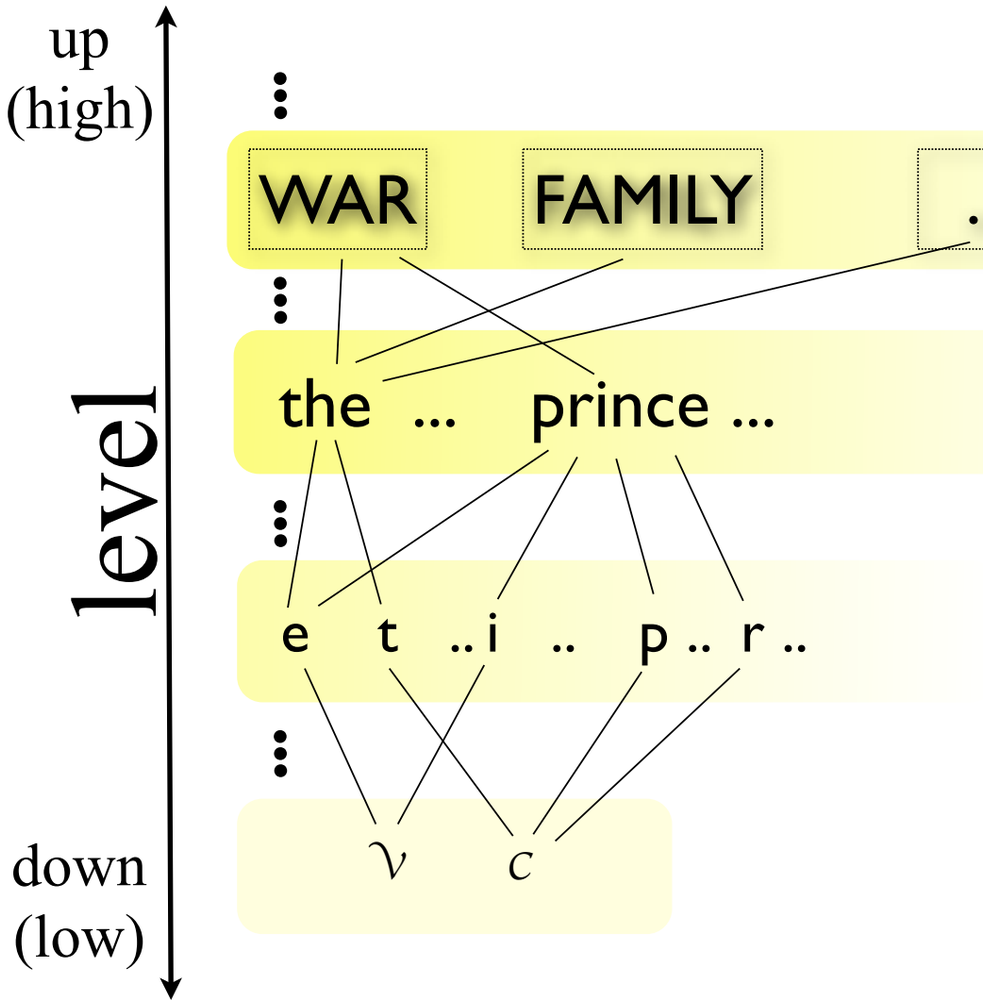
{a,i,o,u}=1000010000001100000000 10...

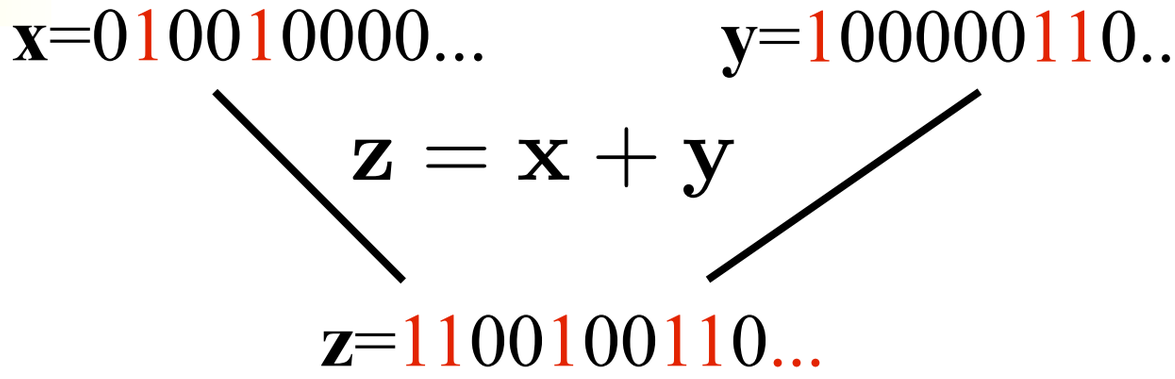
vowels=1100110011011100001100 110...

(+)



Model



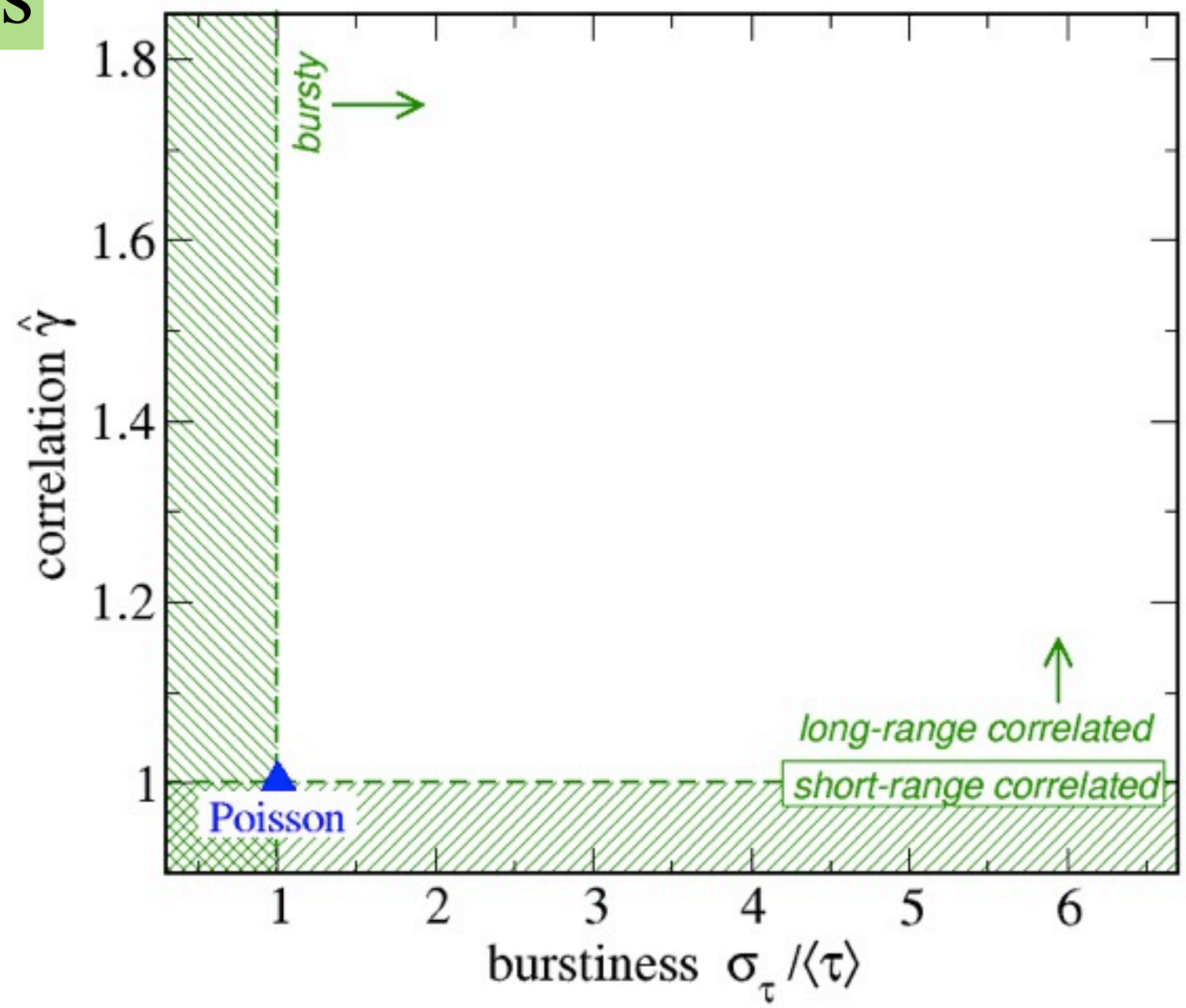


$$\sigma_Z^2(t) = \sigma_X^2(t) + \sigma_Y^2(t) + 2C(X(t), Y(t))$$

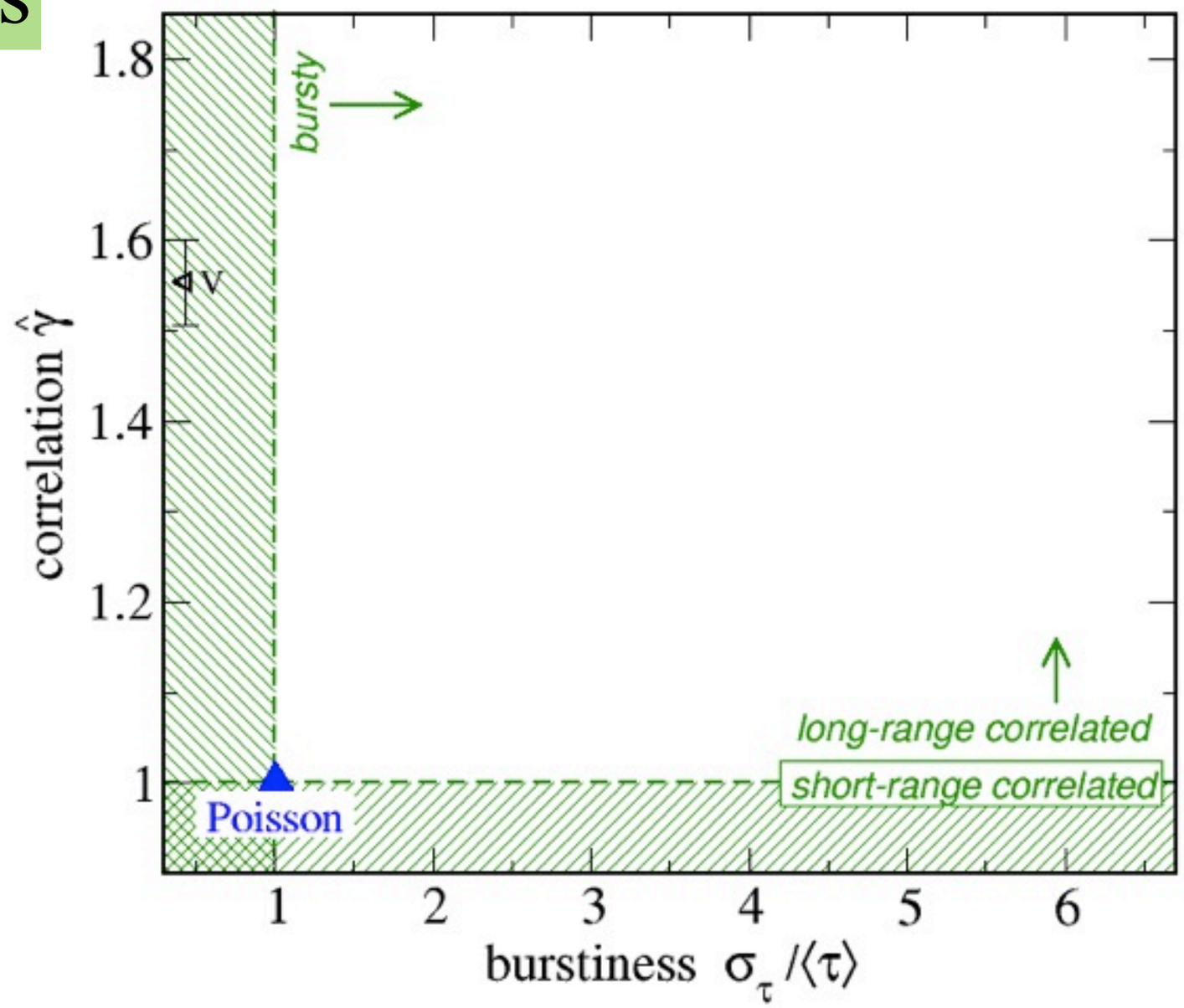
Cauchy-Schwartz inequality: $|C(X(t), Y(t))| \leq \sigma_X(t)\sigma_Y(t)$

	Long-range correlation
Moving up:	necessarily preserved

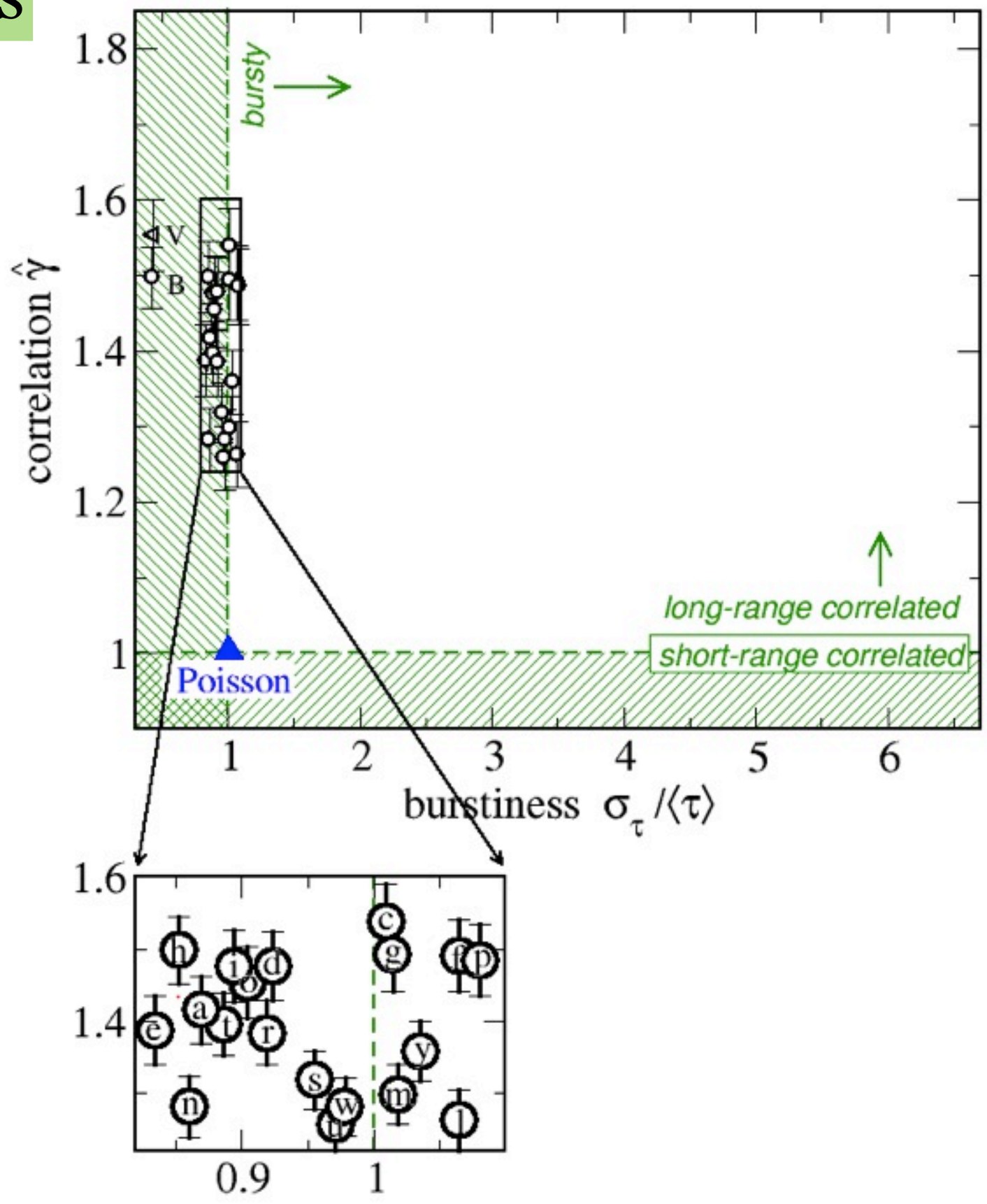
Observations



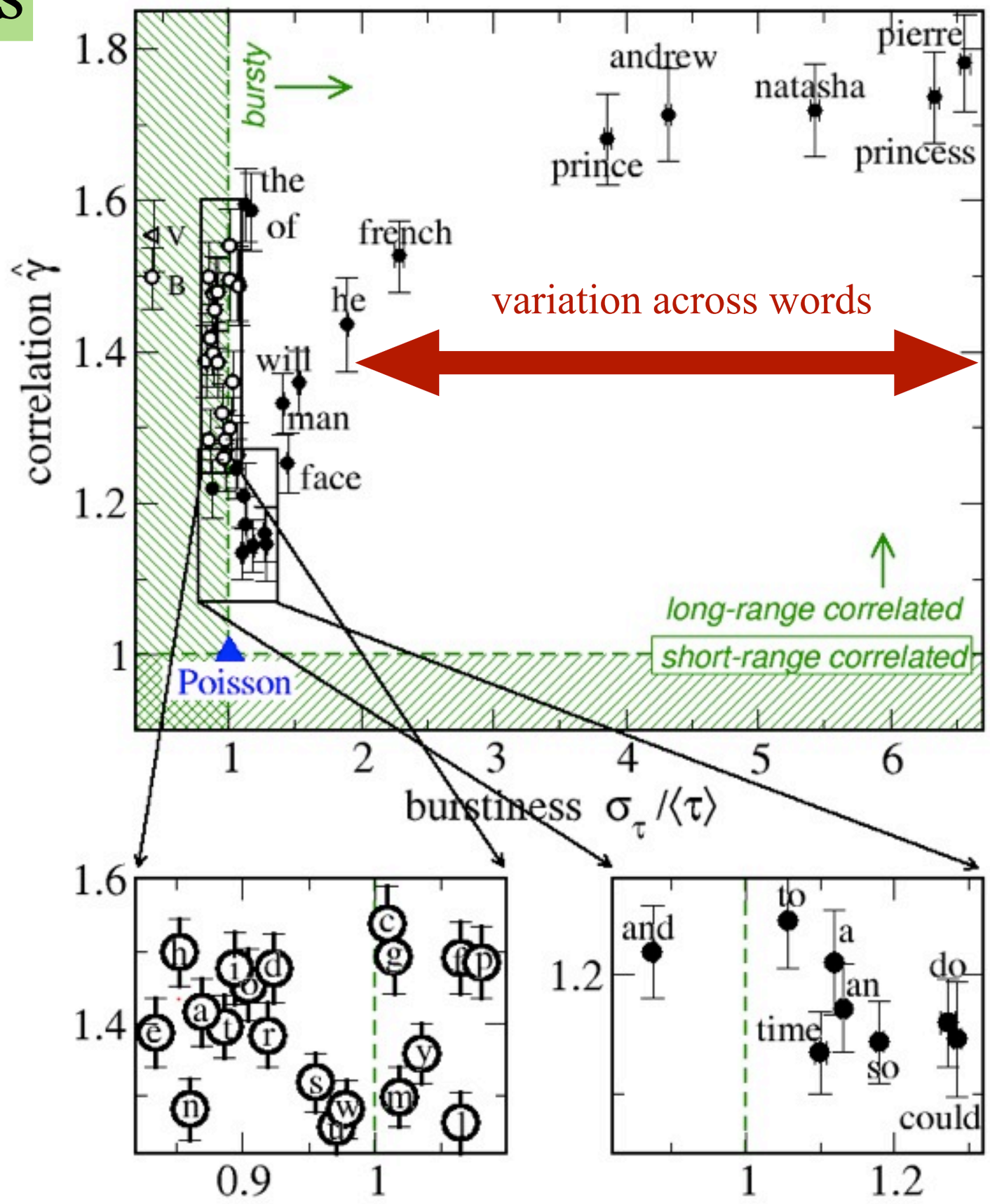
Observations



Observations



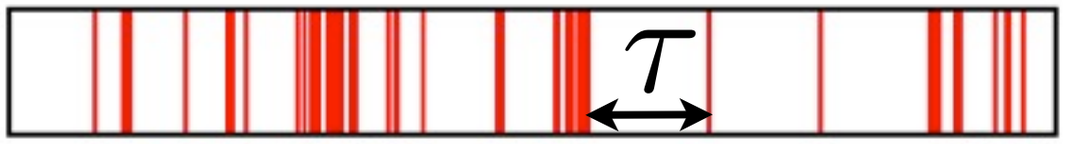
Observations



Burstiness and word types

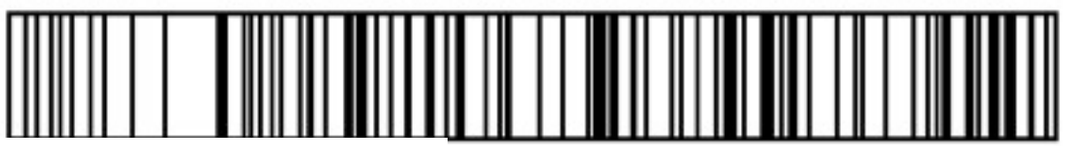
Words with similar frequency: $\langle \tau \rangle \approx 820$

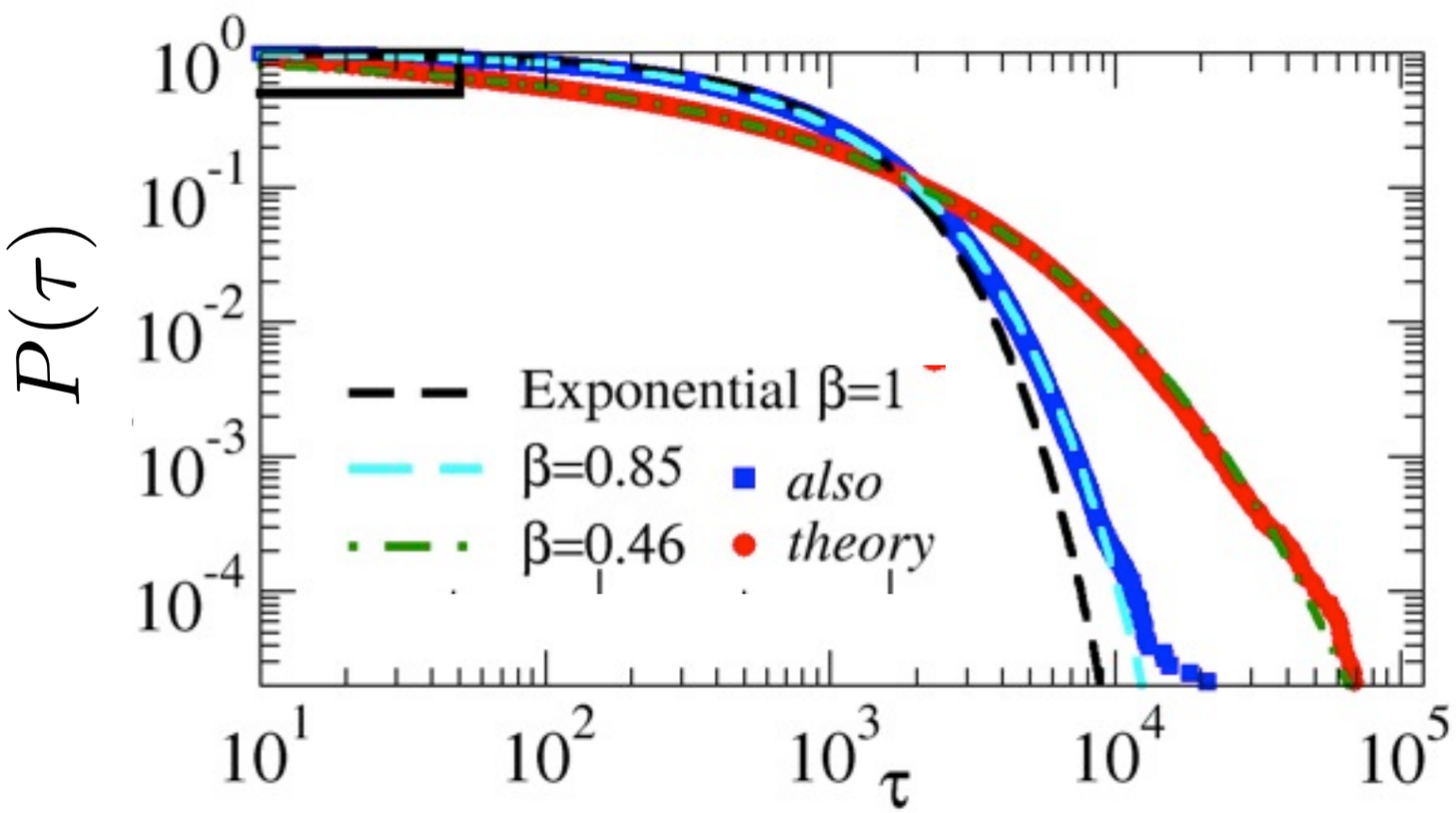
word: theory



word: also

Poisson/random





Weibull distribution

$$P_\beta(\tau) = e^{-a\tau^\beta}$$

E. G. Altmann, J. B. Pierrehumbert, A. E. Motter,"Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words ", PLoS ONE 4 (11) e7678 (2009)

Burstiness and word types

Semantic Classes

Class	Name	Examples of words
1	Entities	Africa, Bible, Darwin
2	Predicates and Relations	blue, die, in, religion
3	Modifiers and Operators	believe, everyone, forty
4	Higher Level Operators	hence, let, supposedly

prince e,t

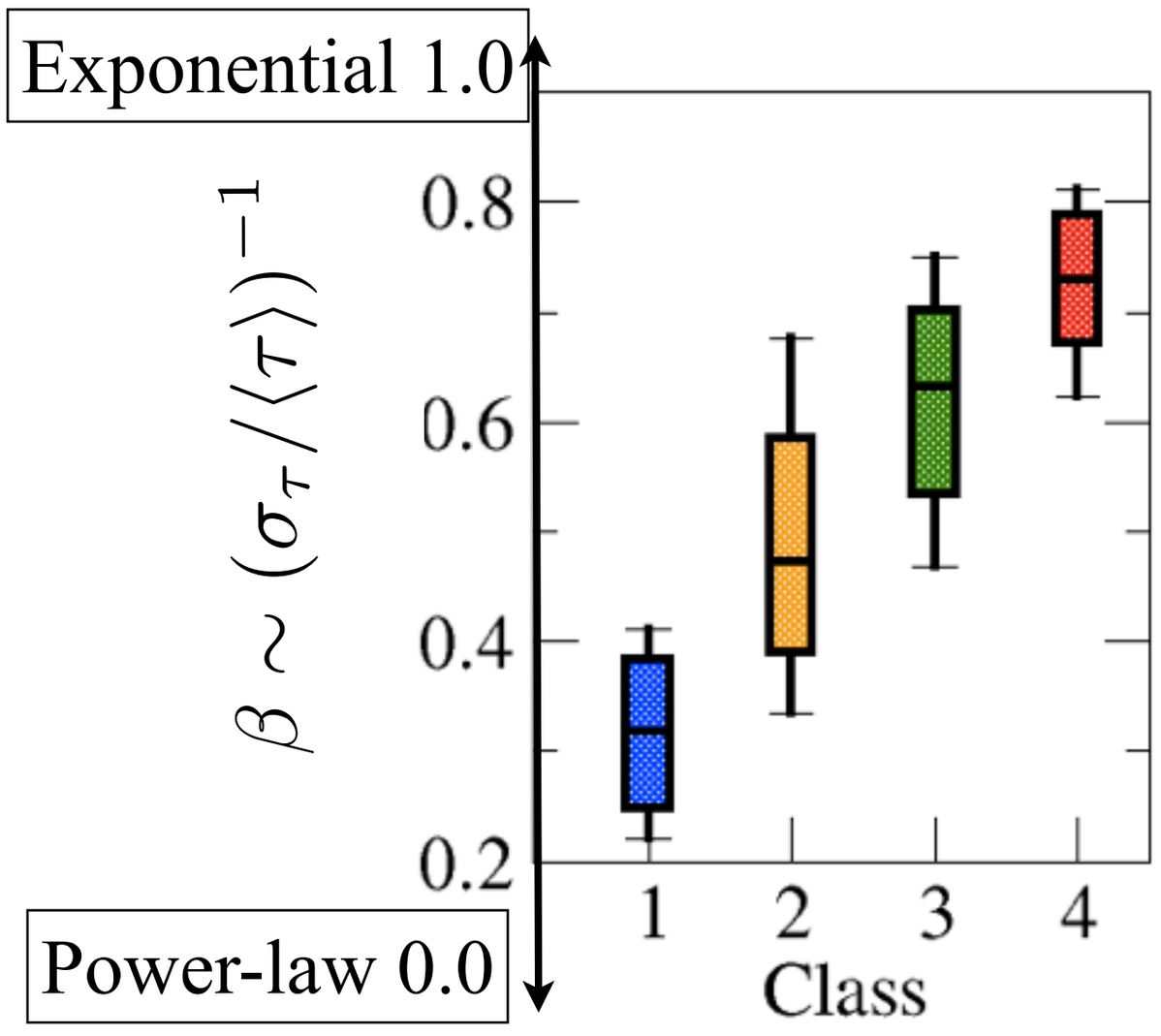
theory <e,t>

also <<e,t>,t>

the <<<>,>>...

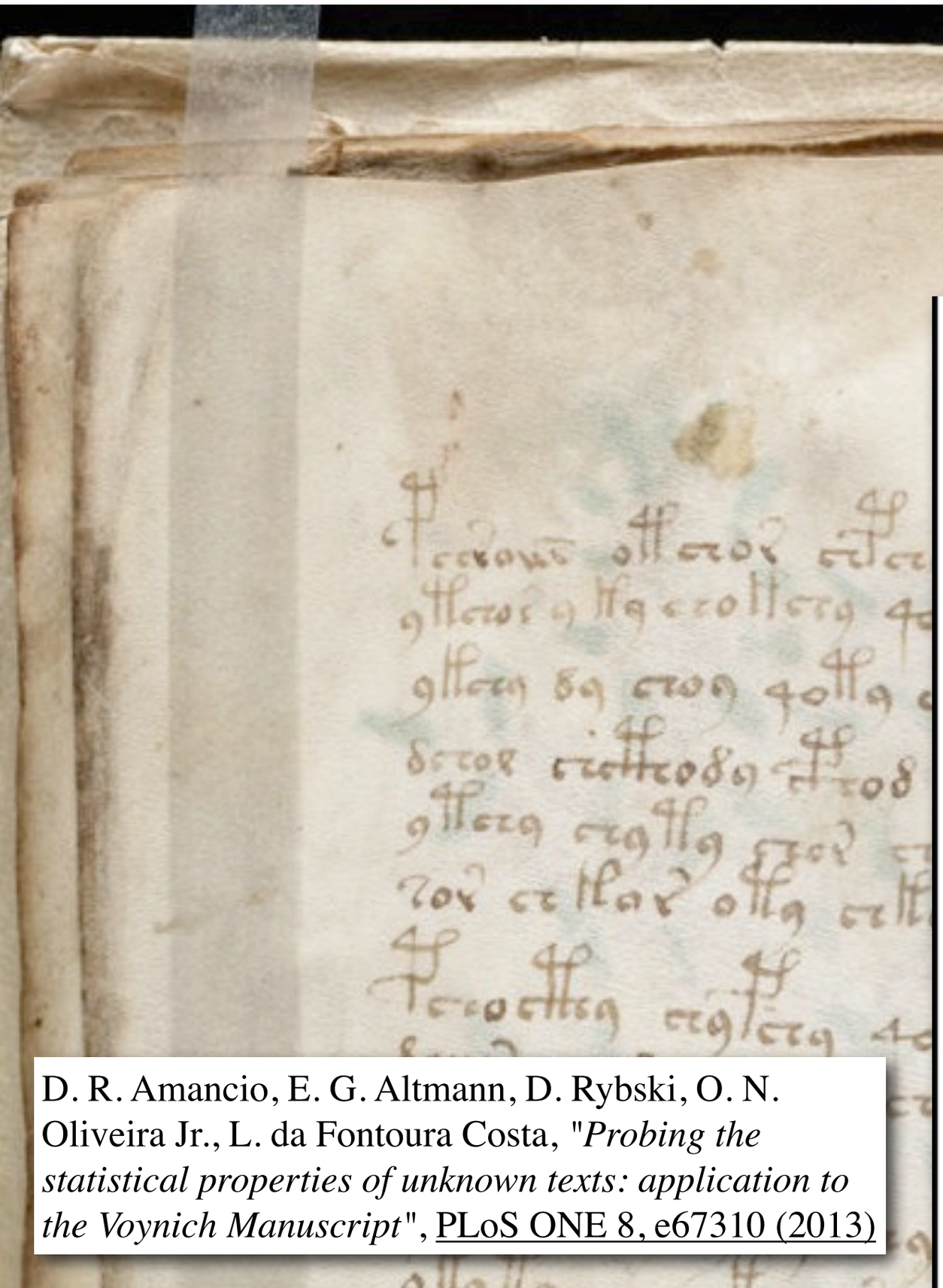
Syntactic (verbs, adjectives, nouns, etc..) relations act on shorter time scales:

$\langle \tau \rangle \gg$ sentence length



Keywords of unknown texts: the Voynich Manuscript

Keywords of unknown texts: the Voynich Manuscript



D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr., L. da Fontoura Costa, "*Probing the statistical properties of unknown texts: application to the Voynich Manuscript*", PLoS ONE 8, e67310 (2013)

Keywords

New Testament

Portuguese	English	German	Voynich
nasceu	begat	zeugete	cthy
Pilatos	Pilates	zentner	qokeedy
céus	talents	himmelreich	shedy
bem-aventurados	loaves	pilatus	qokain
Isabel	Herod	schwert	chor
anjo	tares	Maria	lkaiin
menino	vineyard	Elisabeth	qol
vinha	shall	Etliches	lchedy
sumo	boat	unkraut	sho
sepulcro	demons	euch	qokaiin
joio	five	schiff	olkeedy
Maria	pay	ihn	qokal
portanto	sabbath	weden	qotain
Herodes	hear	heuchler	dchor
talentos	whosoever	tempel	otedy

Summary of conclusions

1. Vocabulary Growth

$$N_{dp}(N_c) = \begin{cases} M, & M \ll M_b, \\ M^{1/\gamma}, & M \gg M_b \end{cases}$$

2. Innovations and Change
   - -accelerated core-vocabulary replacement
   - -quantification of exogenous/endog. factors

3. Text Analysis
   - Burstiness
   - ↓
   - Long-range correlation

Thank you,
for your attention!

* **Martin Gerlach**, José M. Miotto, Fakhteh Ghanbarnejad (MPIPKS, Dresden)
* **Giampaolo Cristadoro**, Mirko Degli Esposti (Univ. Bologna)
* Adilson E. Motter, Janet Pierrehumbert (Northwestern Univ., USA)
* Diego R. Amancio, Osvaldo N. Oliveira Jr., Luciano da Fontoura (USP, Brazil)
* Diego Rybski (PIK, Potsdam)

References:

M. Gerlach, E. G. Altmann, *Stochastic model for the vocabulary growth in natural languages*, Phys. Rev. X 3, 021006 (2013)

D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr., L. da Fontoura Costa, "*Probing the statistical properties of unknown texts: application to the Voynich Manuscript*", PLoS ONE 8, e67310 (2013)

E. G. Altmann, G. Cristadoro, M. Degli Esposti, "*On the origin of long-range correlations in texts*", PNAS 109, 11582 (2012)

E. G. Altmann, J. B. Pierrehumbert, A. E. Motter,"Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words ", PLoS ONE 4 (11) e7678 (2009)