言語データに内在する大域的性質に関して

―統計物理学的観点から―

2017/03/07

東京大学先端科学技術センター
JST さきがけ
田中久美子

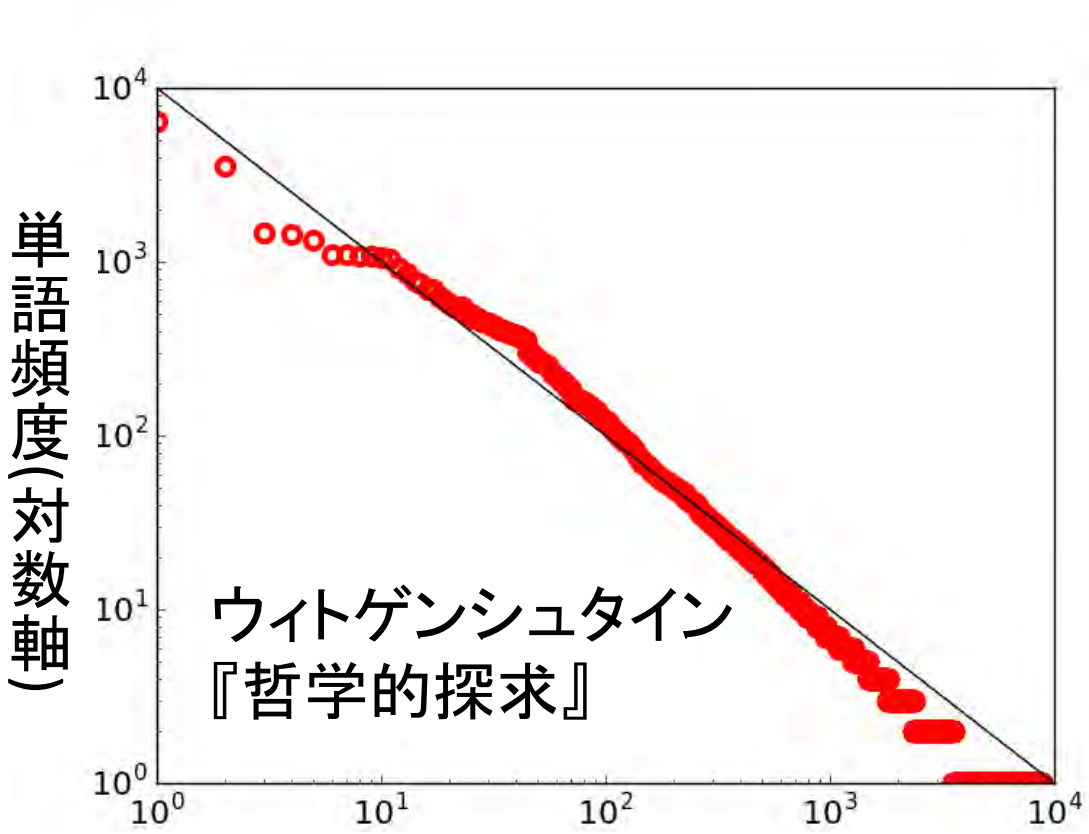
本日の話題

言語に内在する数理的な普遍性/不変性

話題 1　　言語のエントロピーレート
話題 2　　言語の長相関

-物理学誌などに掲載された言語に関する研究
-関連する我々の取り組み
の概要を紹介

Zipf則



単語頻度が多い順に並べた時の順位 (対数軸)

問　　今日の主題

1. このような法則には
   他に何があるのか

2. 法則群の背景にある
   原理は何か

人はZip則を意識的に出そうと思って言語を運用してはいない←言語運用能力に由来?

これまでに議論されてきた考察

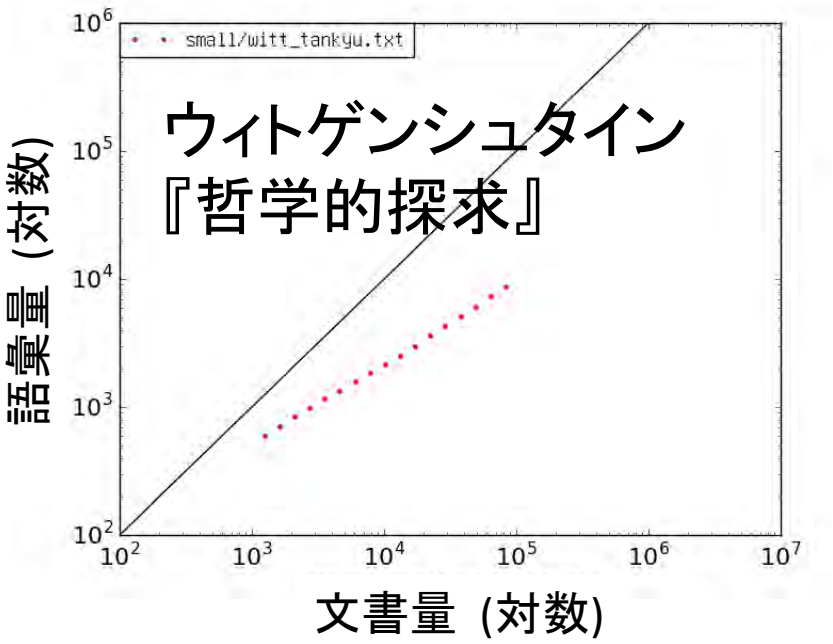
1. エントロピーに関するもの(主に情報理論)　⇐ 話題1
2. 冪則に関するもの(主に複雑系科学)

    Zipf則　順位頻度分布に内在する冪則

    Heaps78/Guiraud54/Herdan64則:　語彙量が文書量に対して冪で増大

    長相関　⇐話題2
3. 他マイナーなもの少数

    ←怪しいものも含まれる



話題1
言語のエントロピーに関する考察

1. YuleのK
   粟飯原俊介氏
2. エントロピーレート
   高比良亮介氏，Lukasz Debowski 准教授との協働研究

G. K. Yuleの提案 (1944)
“Statistical Theory of Literary Vocabulary”

著者推定を行うための良い統計量はあるか。
統計量は文書長に依存する。
「文書長に依存しない統計量」⇦ 文書の保存量のようなもの？

YuleのK

N: 文書の総単語数
$V(m,N)$: m回現れる単語の種類数
C: 適当な定数

$$K = C\left[-\frac{1}{N} + \sum_{m=1}^{m_{\max}} V(m,N)\left(\frac{m}{N}\right)^2\right]$$

すべての単語の相対頻度の2乗和

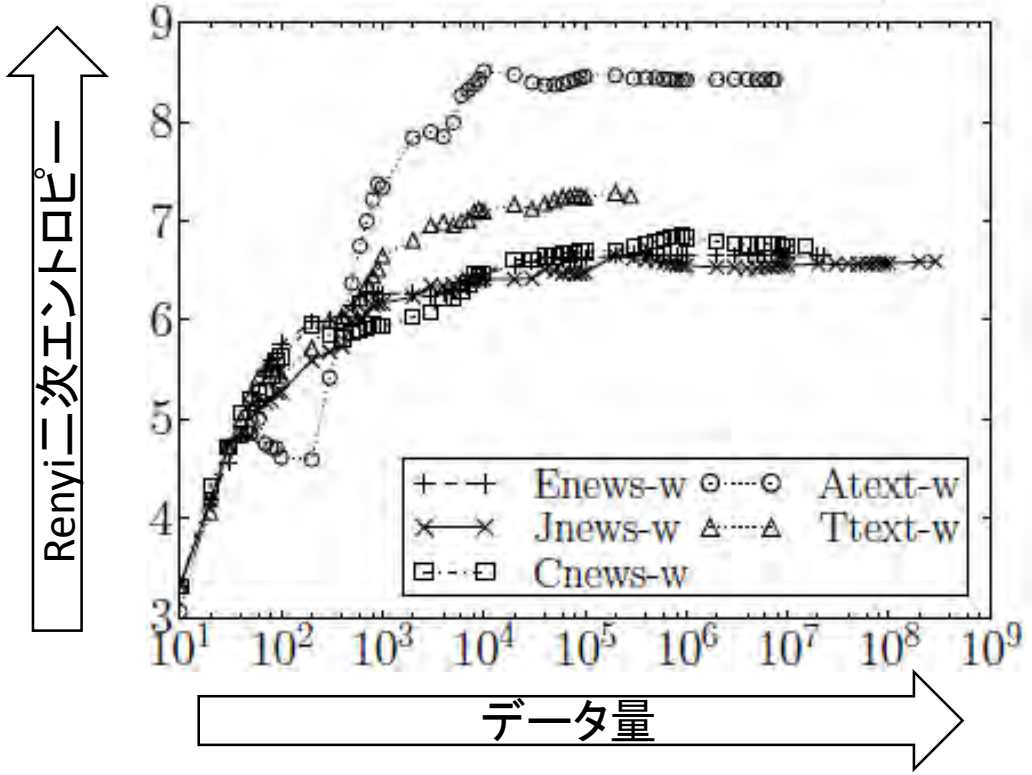
その後のいきさつ:
YuleのKは文書量非依存
只し、著者判別力はない

YuleのKは確率を単語の相対頻度で考えた場合の二次のRényiエントロピーに等価

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_X P^\alpha(X)$$

$\alpha \geq 0, \alpha \neq 1$
X: 要素（文字、単語、それらの列）の集合

- $\alpha = 2$　単語の相対頻度で近似した場合が YuleのKに相当
- $\alpha \to 1$　Shannonエントロピー
  単語相対頻度であれば収束
  言語のエントロピーレートは？
- $\alpha = 0$　Xの種類数。文字の種類数は（ほぼ）有限
  単語の場合は文書量のほぼ対数で増大



相対頻度であれば、シャノンも収束

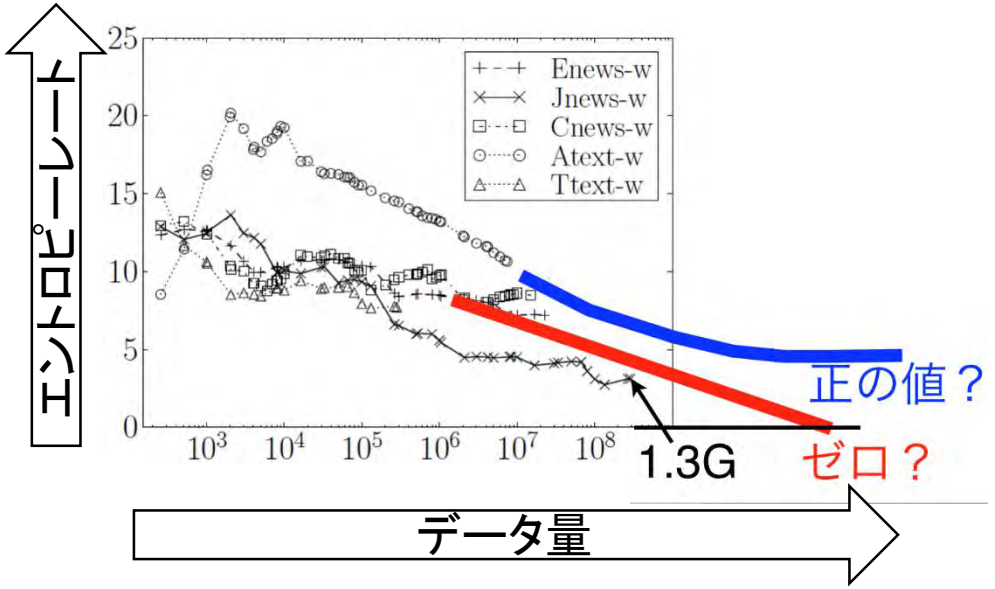
エントロピーレート

時系列 $\boxed{X_1, X_2, X_3, X_4, \ldots, X_n}, \ldots X_N$

X_1^n 長さnのブロック

Block Entropy $\quad H(X_1^n) = -\sum_{X_1^n} P(X_1^n) \log P(X_1^n)$

Entropy Rate $\quad h = \lim_{n\to\infty} \frac{H(X_1^n)}{n}$ 圧縮限界



hを求める研究はShannon以来今日まで続いている

$h > 0$ or $h = 0$ ？

h : 時系列の一要素あたりのbit数
一要素あたり2^hの可能性
長さnの場合2^{nh}の可能性
人間の言語の指数的増大速度

未解決の問い: h >0 なのか？
Hilberg1990はh=0を主張

これまで

1. Entropy Rate一定仮説
   『人間の言語は一定のエントロピーレートをもつ』
   (Genzel & Charniak 2002 ； Levy&Jaeger 2007)
   定式化なし　根拠稀薄

2. 認知実験 (Shannon, Cover&King)
   小規模人数での実験
   - 1.3bpc　英語(Shannon 1951)
   - 1.34 bpc 英語 (Cover & King 1978)

既存研究は
英語での研究が圧倒的
(露、日など他の言語も
報告はされている)

3. 計算機実験 (多くの試み)
   - 1.75 bpc 英語(Brown 1983)　言語モデルを利用
   - 1.45 bpc 英語(Bell, Cleary Witten 1990) 圧縮

   } 無限遠点に補完していない

   他
   過去の研究では英語が主

以下、データを圧縮し、補完してみた

ユニバーサル符号

$$h = \lim_{n\to\infty} \frac{H(X_1^n)}{n}$$

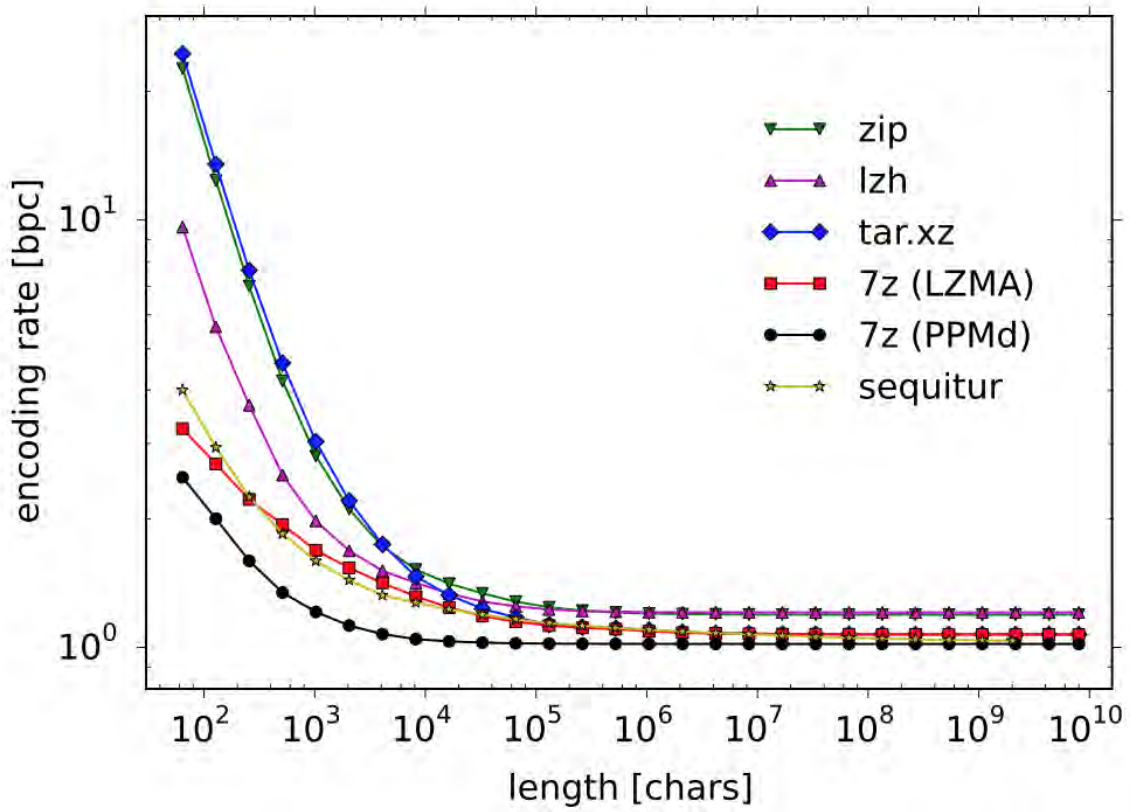
$R(X_1^n)$ 圧縮した文書のbit数

$r(n) \equiv R(X_1^n)/n$ 符号化レート > h

ユニバーサル符号

$$\lim_{n\to\infty} r(n) \rightarrow h$$

但し、時系列に定常性,エルゴード性を仮定



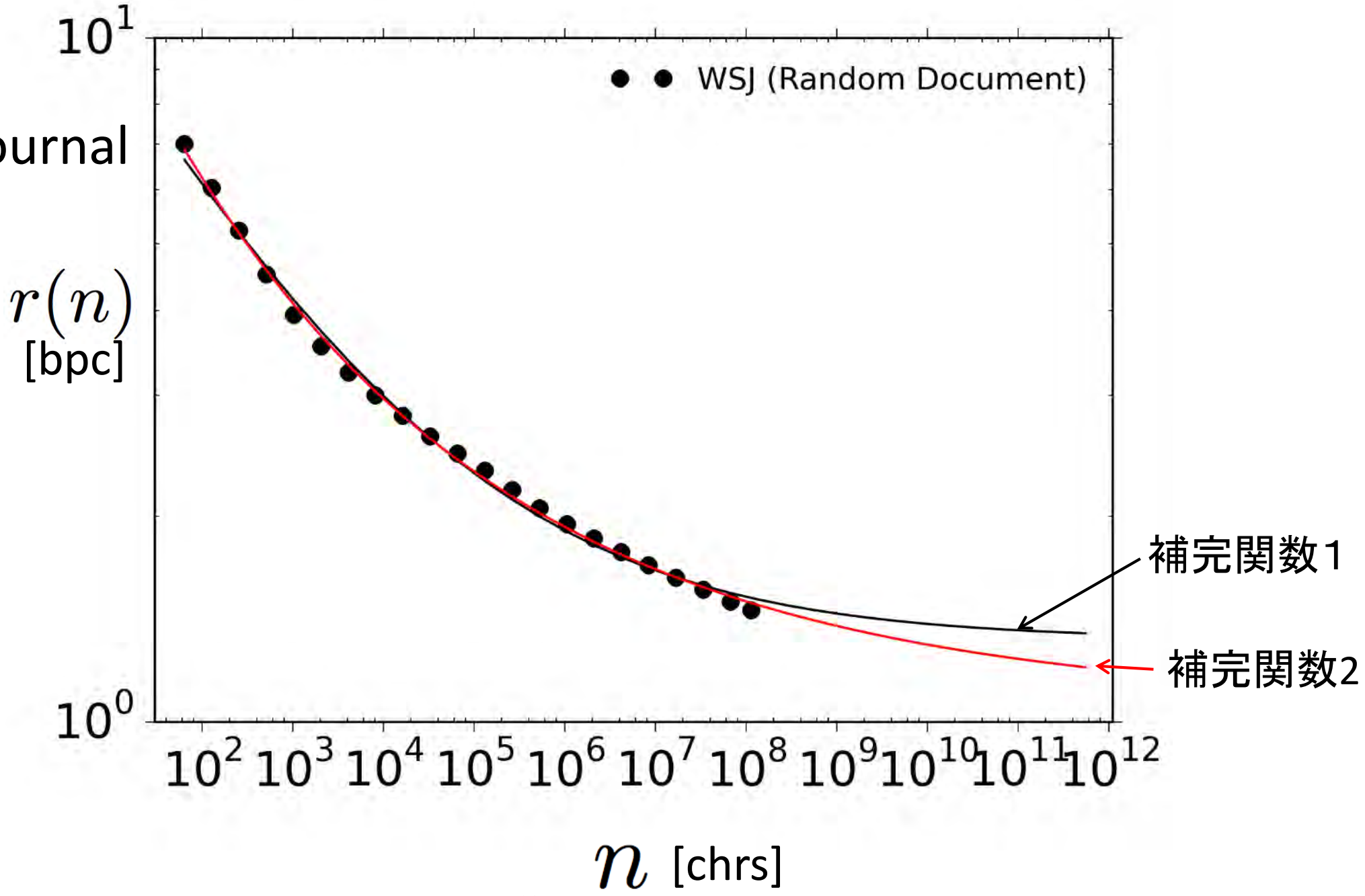
今日使われる圧縮器はどれも理論的にはユニバーサル

収束の速さが異なる

乱数を用いて理論値に収束したPPMを利用

圧縮例

Wall Street Journal



無限遠点を得る補完関数　ansatzに過ぎない

$0 < \beta < 1$

- Hilberg 1990

$f_1(n) = An^{\beta-1} + h$

- Ebeling & Nicholis 1991 , Schumann&Grassberger 1996

$f_2(n) = An^{\beta-1} \ln n + h$

- 今回使ってみたもの ← データによりfit

$f_3(n) = \exp(f_1(n))$

β　曲線の落ち方、学習のしやすさ, 予測の改善率

h　言語のもつ本来的な乱雑さ

データ

- 6言語 18コーパス
  (英, 仏, 露, 中, 日, 韓)
- 新聞, 文学作品
- 日中はアルファベットに変換したものも用意(ローマ字, pinyin)
- 1コーパスの最大量は7.8G
- 英語の総量 30G

Text	Language	(c
	Large Scale Random	
Agence France-Presse	English	409600
Associated Press Worldstream	English	652427
Los Angeles Times/Washington Post	English	154523
New York Times	English	782787
Washington Post/Bloomberg	English	9741
Xinhua News Agency	English	192988
Wall Street Journal	English	11286
Central News Agency of Taiwan	Chinese	67818
Xinhua News Agency of Beijing	Chinese	38383
People's Daily (1991-95)	Chinese	10150
Mainichi	Japanese	84760
Le Monde	French	72734
KAIST Raw Corpus	Korean	13087
Mainichi (Romanized)	Japanese	191610
People's Daily (pinyin)	Chinese	24755
	Small Scale	
Ulysses (by James Joyce)	English	151
À la recherche du temps perdu (by Marcel Proust)	French	725
The Brothers Karamazov (by Fyodor Dostoyevskiy)	Russian	182
Daibosatsu toge (by Nakazato Kaizan)	Japanese	454

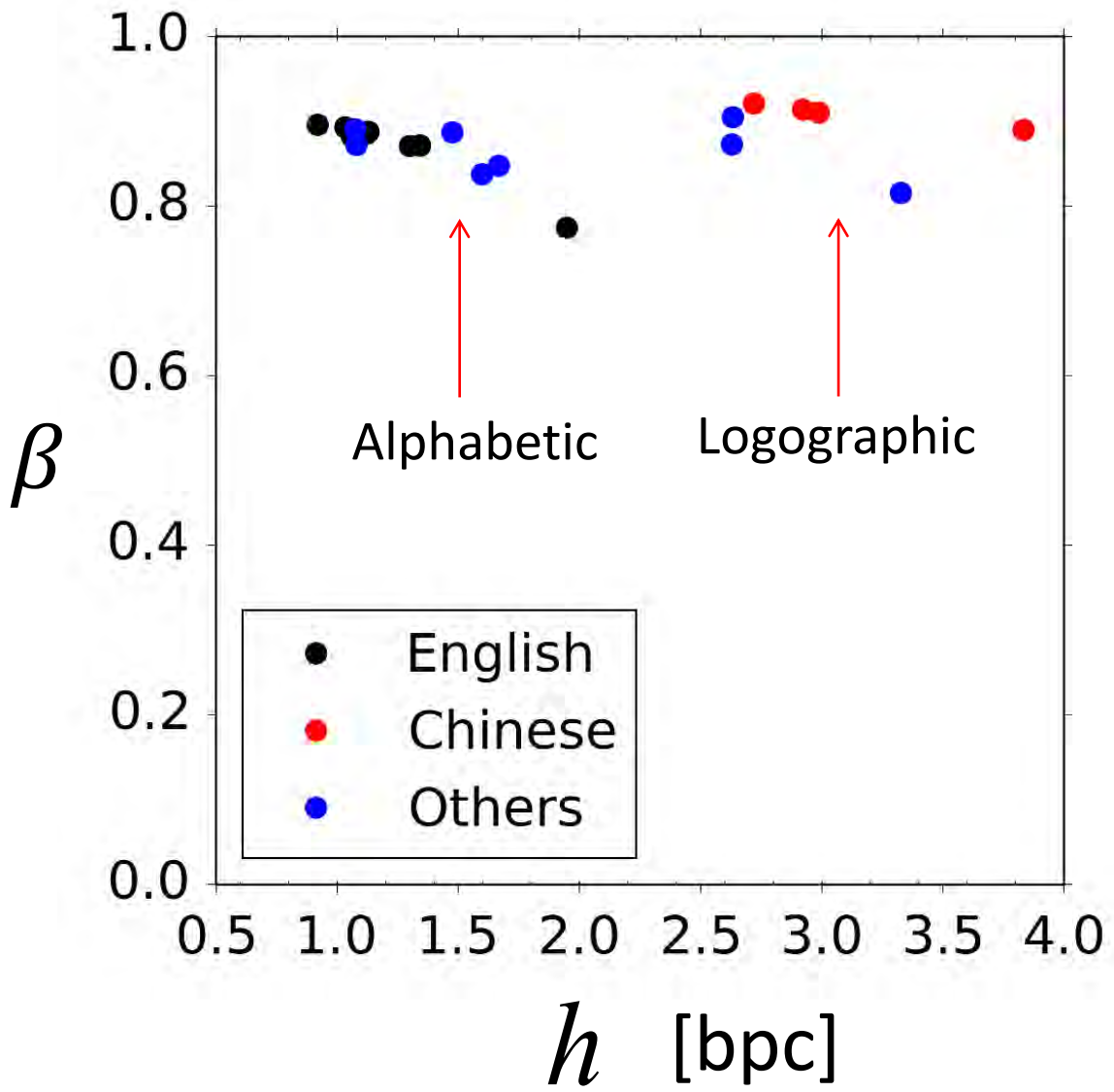
$h-\beta$上で全てのデータをプロット

以下二つのことが言える

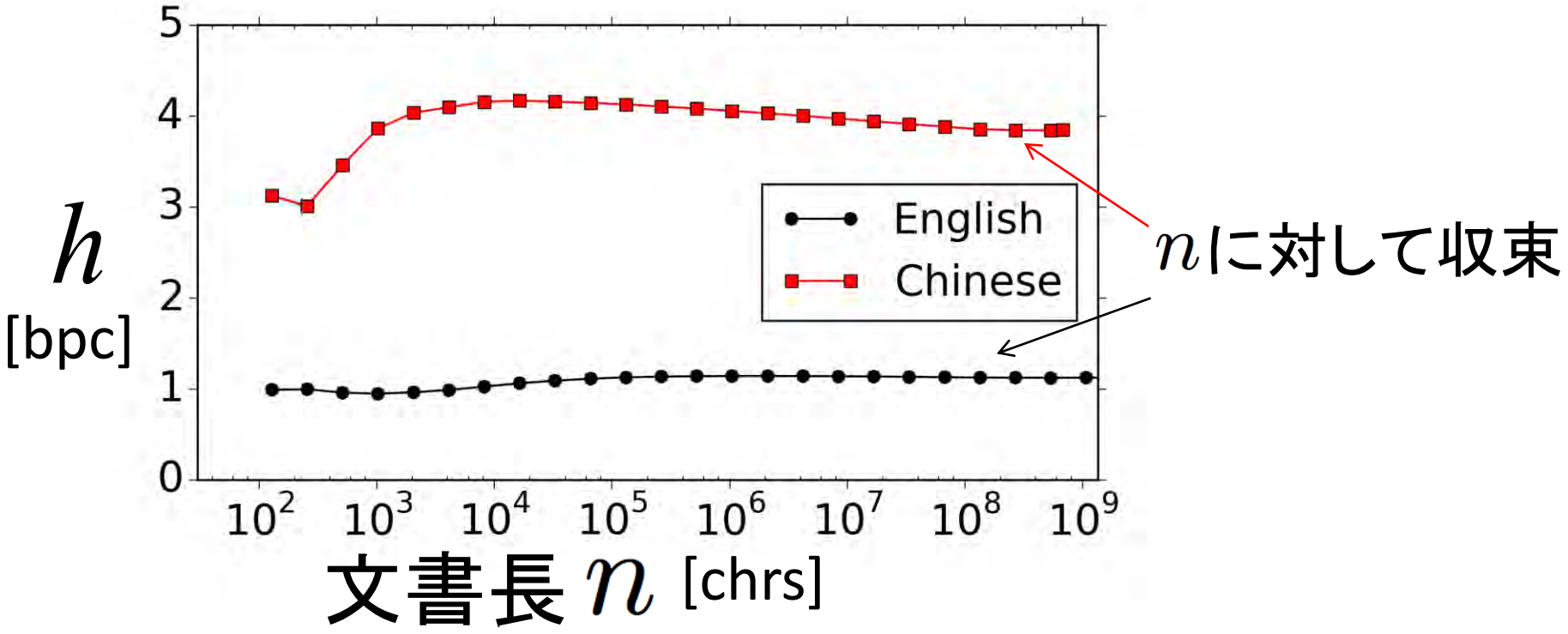
1. エントロピーレートは、文字種に依存
   ざっと表音か表意か

2. βは言語を問わず似た値
   -$f_3(n)$の場合、0.884, 標準偏差 0.034
   -Universalな変量
   -PPMによる自然言語の学習難易度は
   変わらない

$f_1(n)$でも同様の結果



hのデータ量依存性：保存量っぽい？



話題1のまとめ

言語のエントロピーにまつわる研究について近況を紹介
-YuleのK, 推定されたエントロピーレートはデータ量不変
　保存量のようなものに喩えられる？
-自然言語の学習しやすさ、乱雑さ(β)は普遍の値

結局、自然言語のエントロピーレートは正か？
-上限は圧縮を利用すると収束し、正
-理論的な事柄はいずれも定常性、エルゴード性が前提
　　⇒ 究極的には何も言えていない

-得られた普遍量としてのhやβは何であるのか
-問は別の何かでなければならないのではないか

話題2
言語の長相関に関する考察

Giessen大学理論物理学研究所
Armin Bunde教授
との協働研究

言語はフラクタル？

- フランス語の音韻構造はフラクタル?
- 単語の出現はフラクタル?
  ホメロスのイーリアスの中の稀な単語は、固まってあらわれる(van de Boas2004)

言語で何がどの程度本当にフラクタル？　以下、『長相関』を特に考える

長相関に類する解析の方法論は数値時系列に対するものばかり。。。。。

A. 自己相関関数
B. Fluctuation Analysis (以下FA, 発展手法としてDFA, MF-DFA)
C. Hurst解析

言語は非数値時系列

言語を数値データ変換する問題

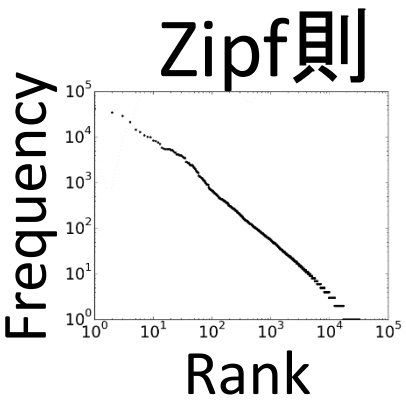
- バイナリ列に変換 (1995: Ebeling and Neiman ; 2012 Altmann et al.)

  文字・単語・品詞を限定し、その出現位置のみ1, それ以外は0

	Oh	Romeo	,	Romeo	,	wherefore	art	thou	Romeo	?
Romeoの場合	0	1	0	1	0	0	0	0	1	0

- 単語のrankの列に変換 (2002 Montemurro and Pury)

Oh	Romeo	,	Romeo	,	wherefore	art	thou	Romeo	?
1234	2435	23	2435	23	3245	845	354	2435	54



- 単語長の列に変換

Oh	Romeo	,	Romeo	,	wherefore	art	thou	Romeo	?
2	5	1	5	1	9	3	4	5	1

既存研究:『言語には長相関がある』
やり方に問題が大きい

1. 言語の数値データへの変換が恣意的にすぎる
   その結果のHurst 指数が0.XXと提示されても、何とも解釈が難しい

2. 手法の問題
   -用いた手法の弱点をふまえていない(Hurst手法, FA)
   -最新の成果が使われていない(DFA, MF-DFA)
   -結果の解釈がいい加減である(cross-overなど難しい現象)

3. 稀な単語が低頻度にすぎ、長相関を捉えきれていない

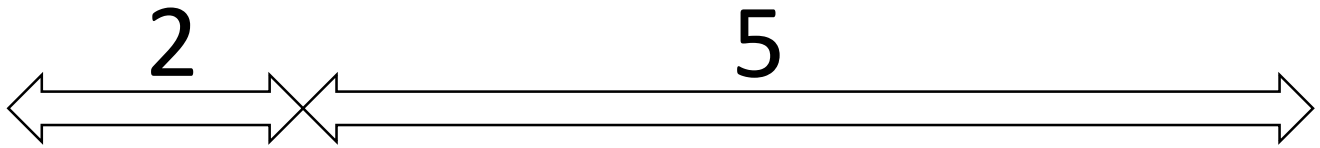
もう少しきちんと捉える手法はないだろうか。

間隔時系列＋極値解析

1. 言語の間隔を時系列で捉える　→ 語の恣意性と無縁に

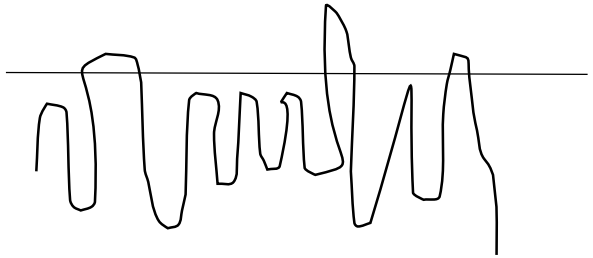
Oh Romeo, Romeo, wherefore art thou Romeo?

For “Romeo”　0　1　0　1　0　0　0　0　1　0



2. 稀な単語は出現数が少なすぎる→極値解析

地震、気象etc.

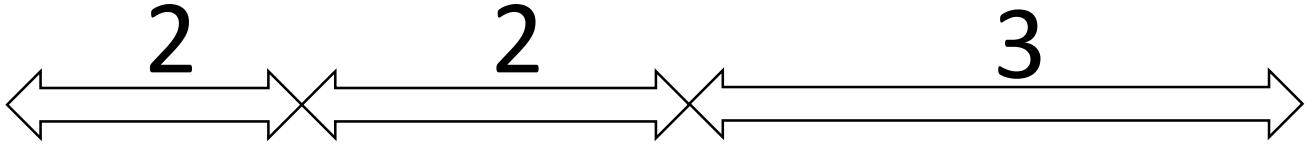


ある閾値を超える稀現象を十把一絡げに考える

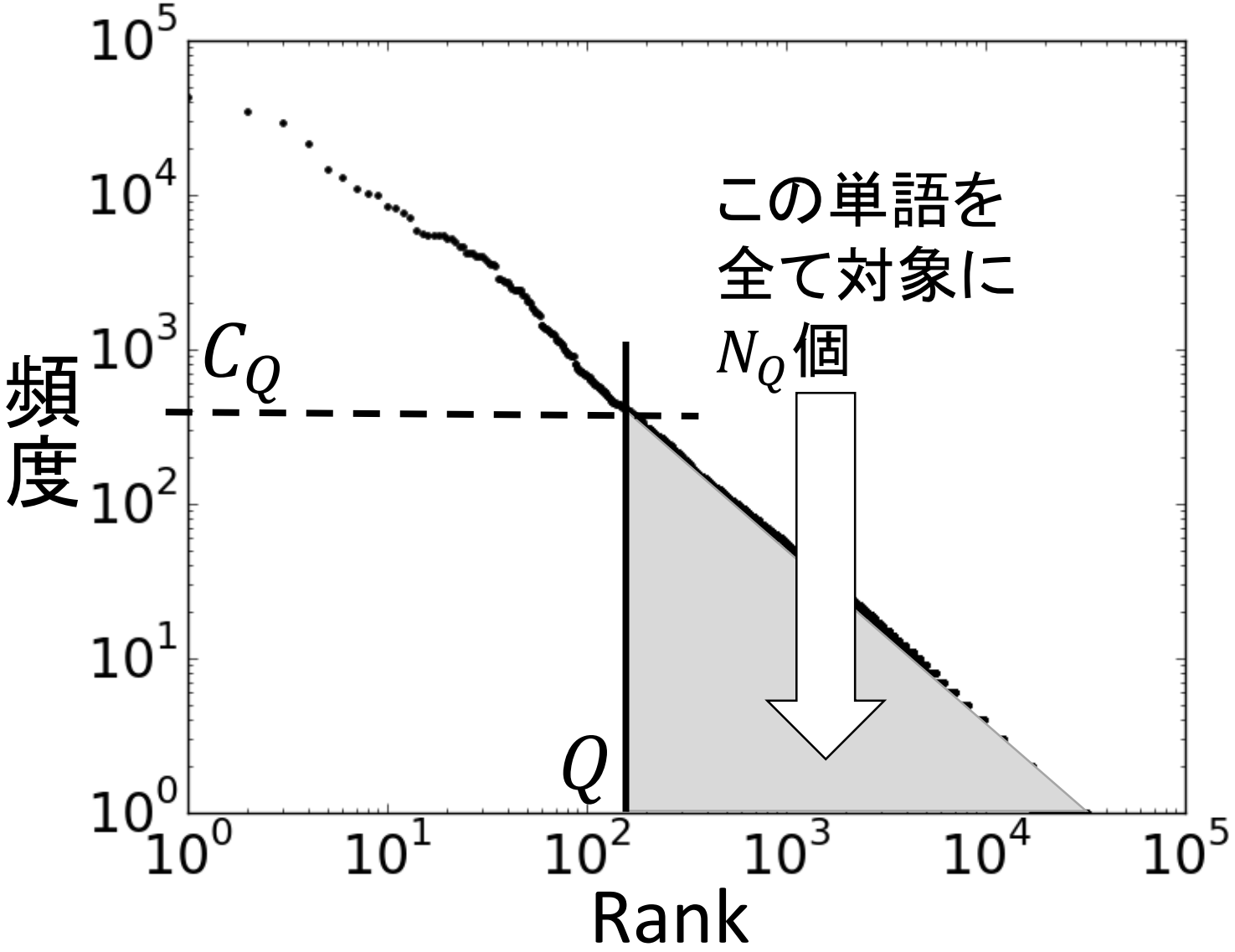
⇒言語：ある閾値以下の稀な単語を十把一絡げに考える

Oh Romeo, Romeo, wherefore art thou Romeo?

{Romeo, wherefore}　0　1　0　1　0　1　0　0　1　0



順位頻度分布上では

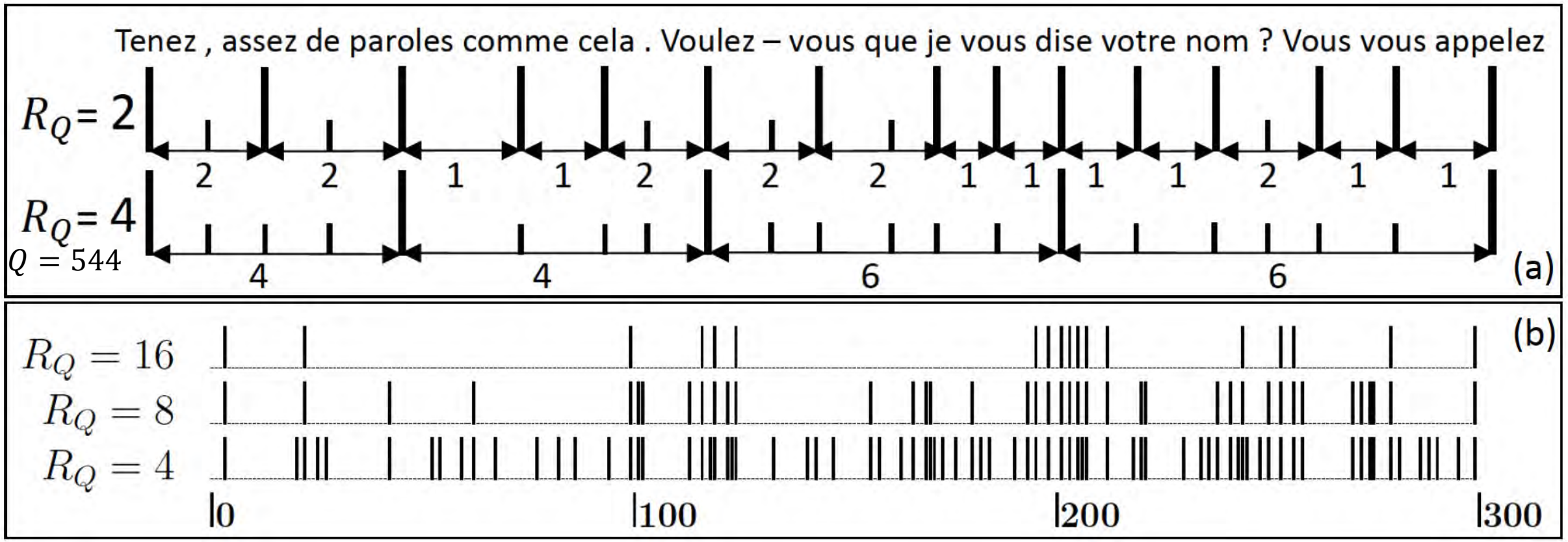


N　総単語数

平均間隔長

$$R_Q \approx \frac{N}{N_Q}$$

レミゼラブルに対する図示

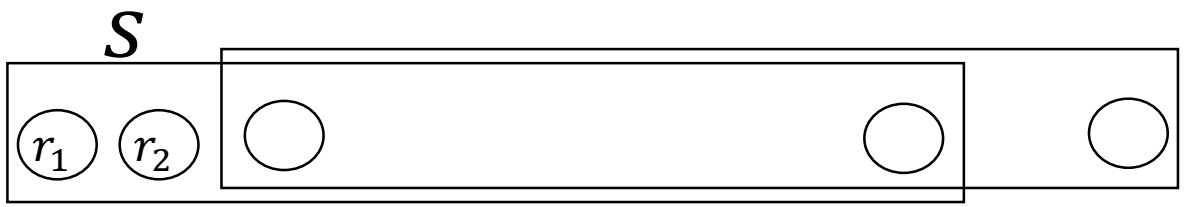


稀な単語は確かにまとまって現れている

間隔時系列の数理モデル

“Les Miserables” for $R_Q = 16$

r_i：間隔，R_Q：平均



間隔分布の振る舞い

超過関数

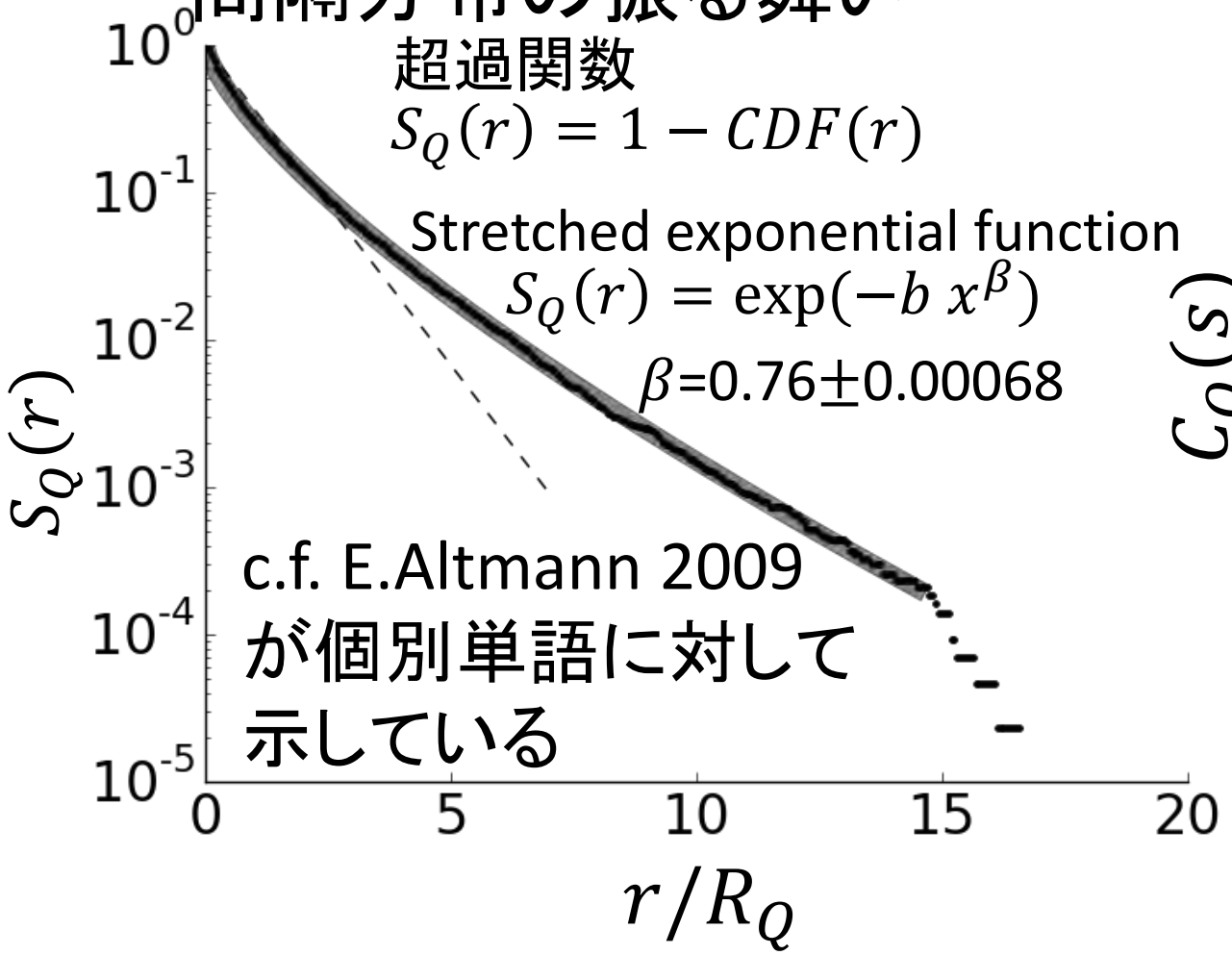
$S_Q(r) = 1 - CDF(r)$

Stretched exponential function

$S_Q(r) = \exp(-b\, x^{\beta})$

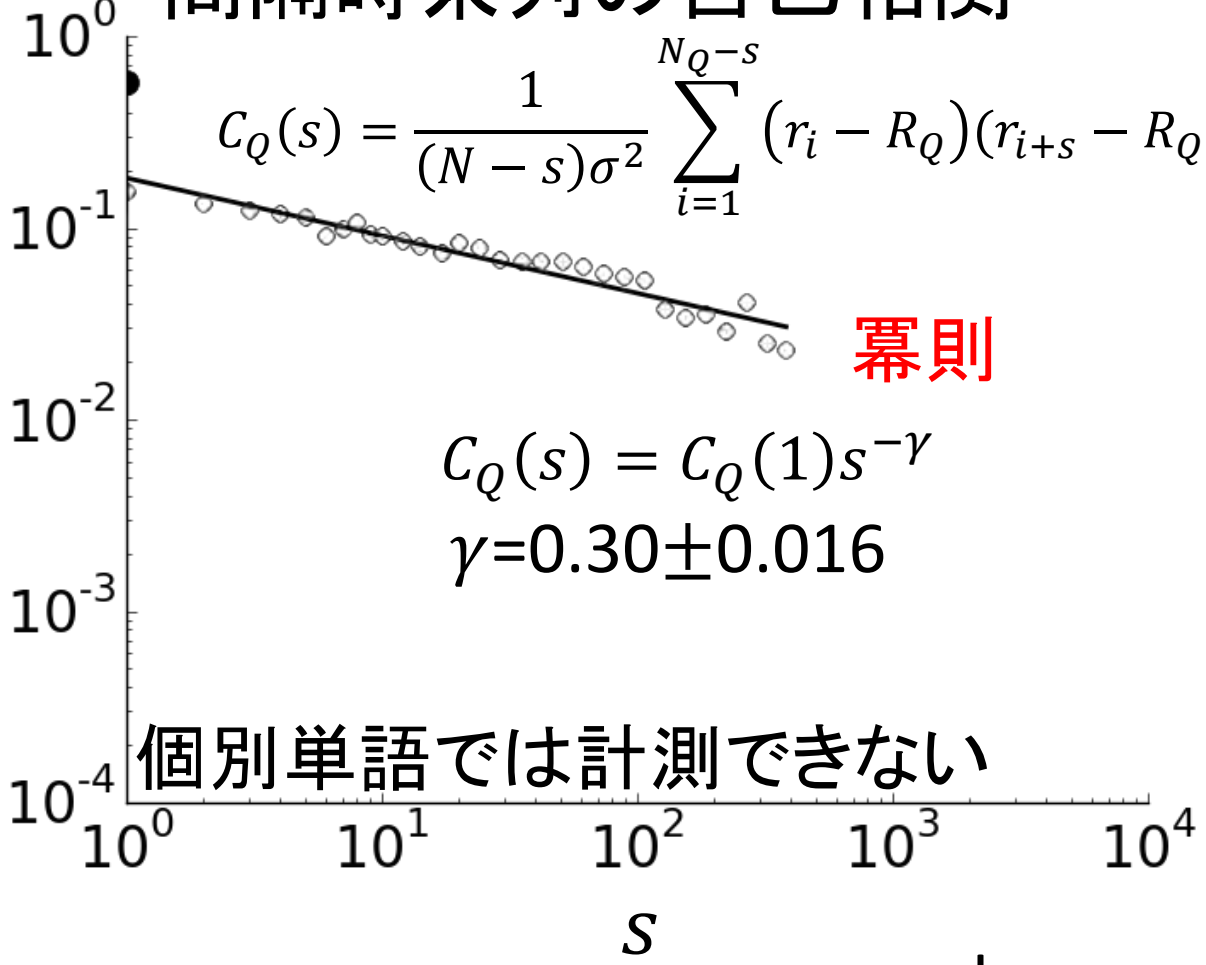
β=0.76±0.00068

c.f. E.Altmann 2009
が個別単語に対して
示している



間隔時系列の自己相関

$$C_Q(s) = \frac{1}{(N-s)\sigma^2} \sum_{i=1}^{N_Q - s} (r_i - R_Q)(r_{i+s} - R_Q)$$



冪則

$C_Q(s) = C_Q(1)s^{-\gamma}$

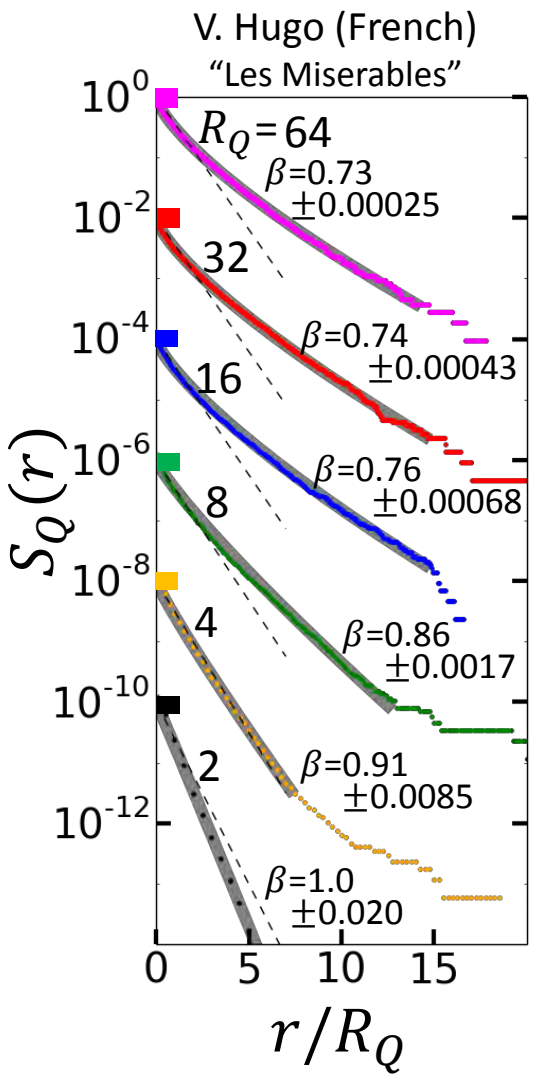
γ=0.30±0.016

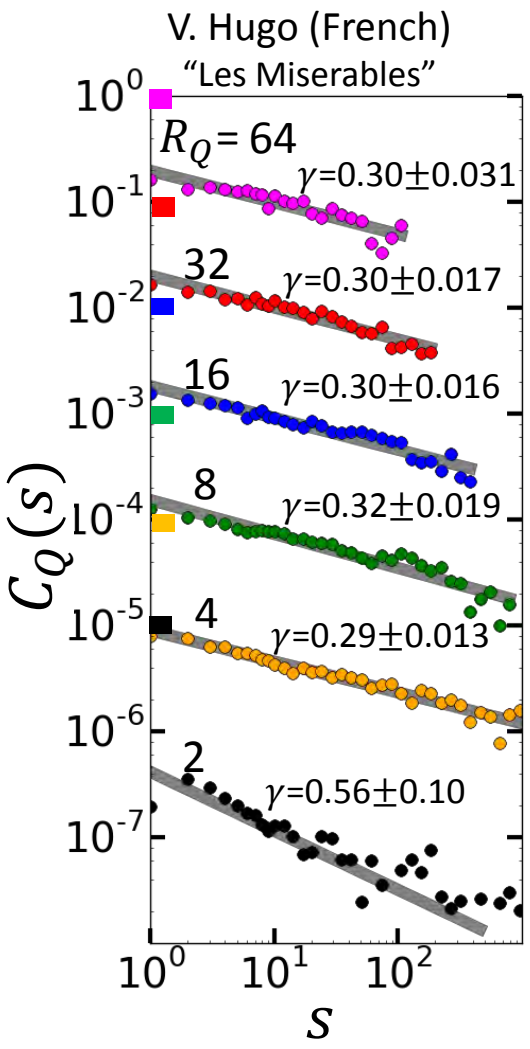
個別単語では計測できない

s: lag

稀度を変化: レ・ミゼラブル

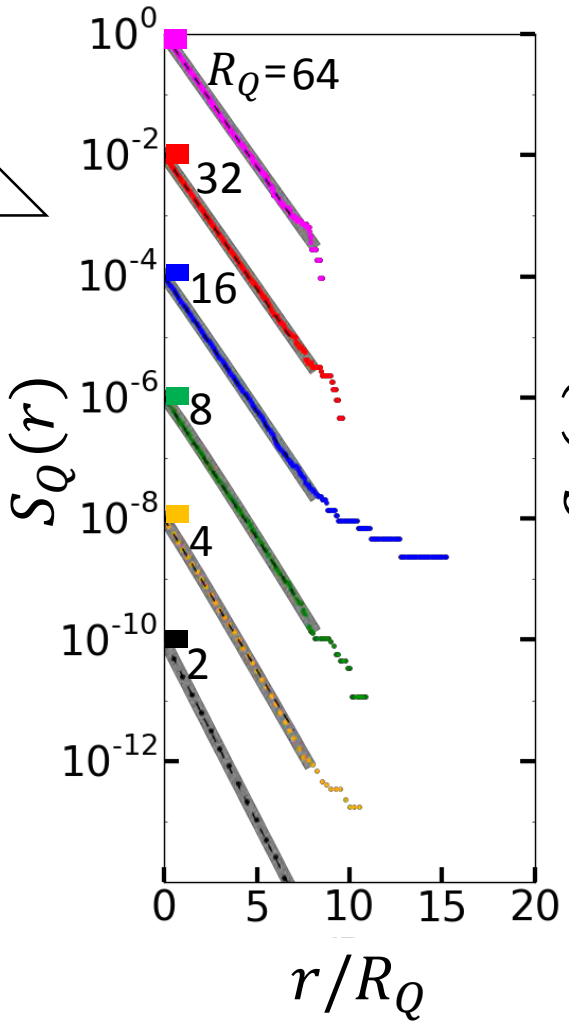
$R_Q = 2,4,8,16,32,64$

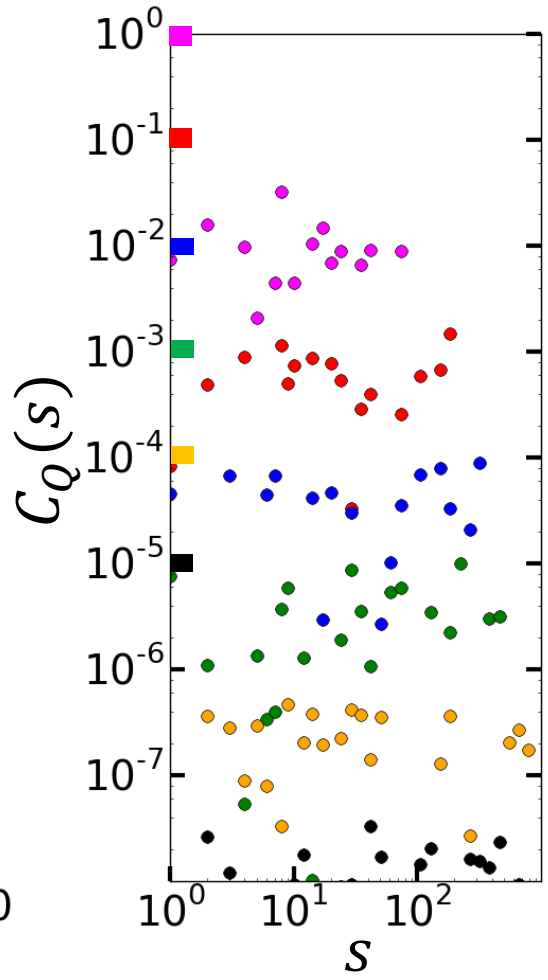




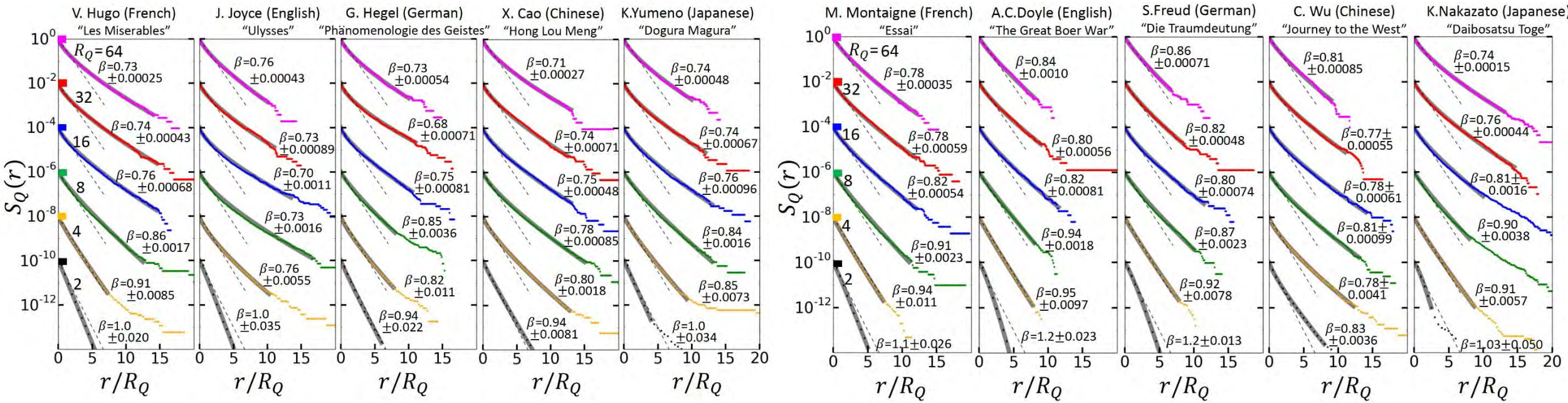
より稀

単語シャッフルしたレ・ミゼラブル





10編の単著に対する$S_Q(r)$ $R_Q = 2,4,8,16,32,64$

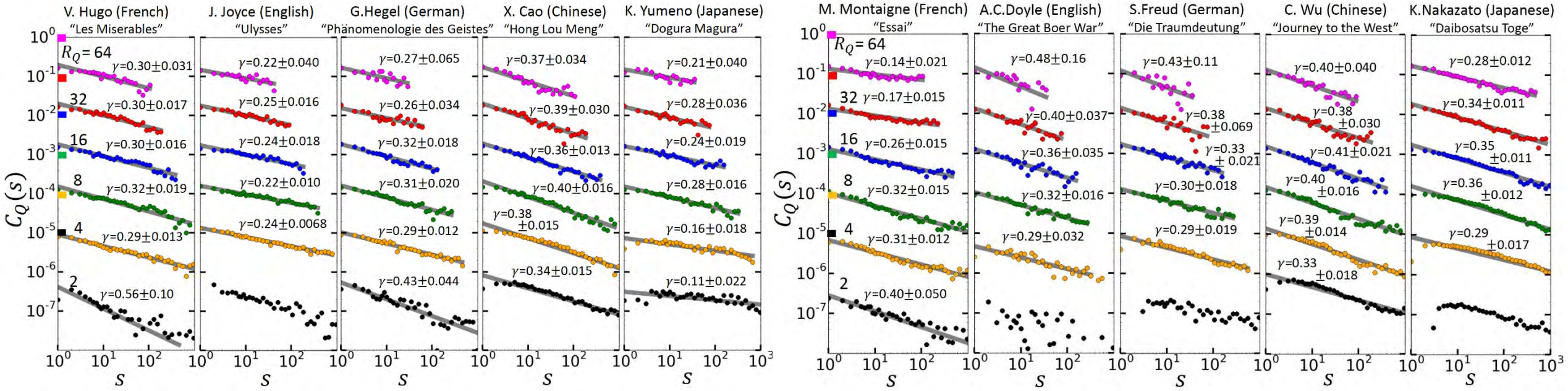


大きなRQではβの値はより安定

$R_Q = 2$	平均1.1	標準偏差 0.13
$R_Q = 4$	0.86	0.067
$R_Q = 8$	0.85	0.059
$R_Q = 16$	0.77	0.037
$R_Q = 32$	0.76	0.037
$R_Q = 64$	0.77	0.048

10編の単語(5言語)で傾向は一致している
ジャンルも小説、随筆、哲学的考察、
など多岐に渡る

10編の単著に対する $C_Q(s)$ $R_Q = 2,4,8,16,32,64$

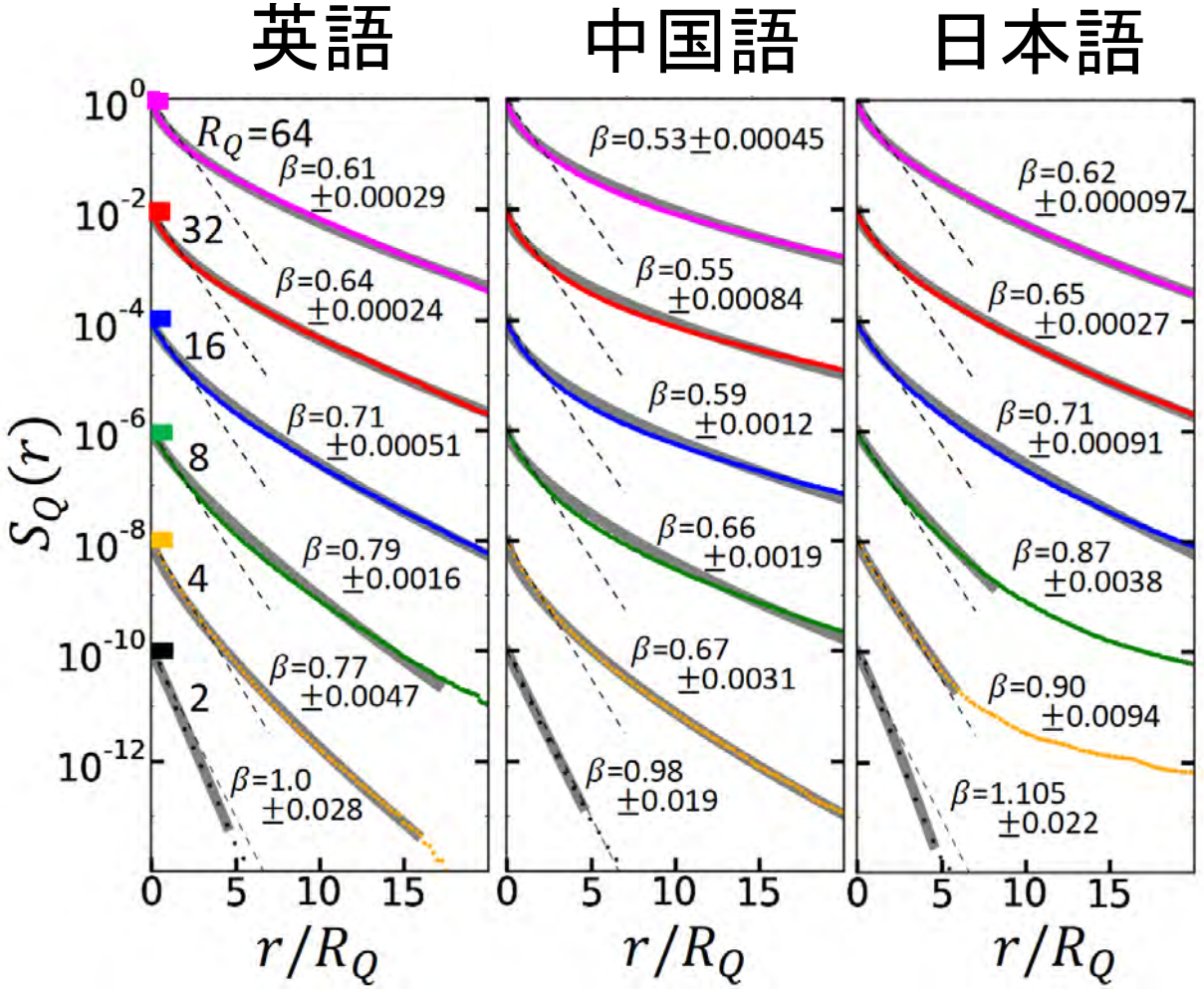


大きいR_Qに対しては指数は安定

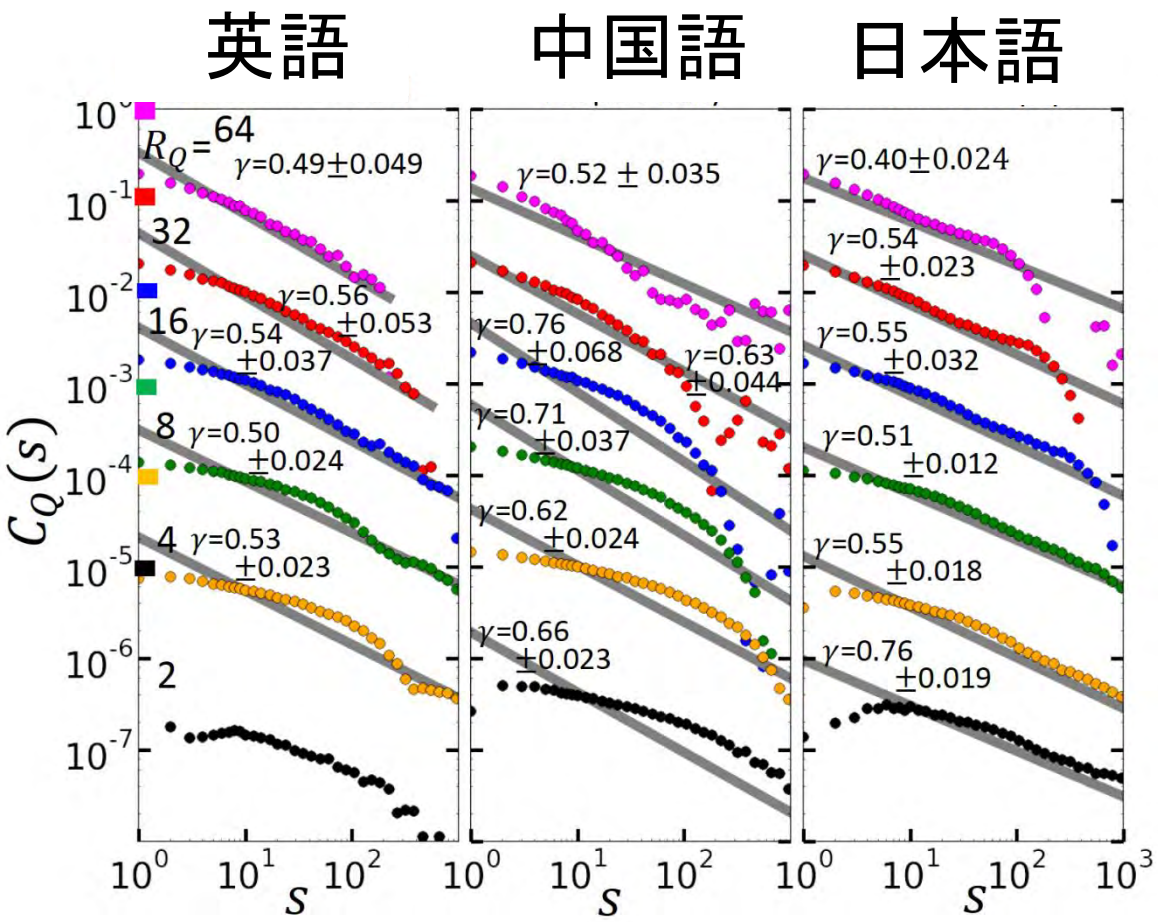
	平均	標準偏差
$R_Q = 2$	0.36	0.036
$R_Q = 4$	0.31	0.040
$R_Q = 8$	0.34	0.035
$R_Q = 16$	0.34	0.049
$R_Q = 32$	0.33	0.084
$R_Q = 64$	0.35	0.012

10編の単語(5言語)で傾向は一致している
ジャンルも小説、随筆、哲学的考察、
など多岐に渡る

新聞(複数著者)　100M　から 1.3 Gbytesの日にち順データ



指数からより乖離



やや異なる様相

話題2 のまとめ

- 自然言語には単語の現れ方(間隔)には長相関がある(単著の場合)
- 解析の方法を改良: 間隔＋ 極値解析を行うと安定した解析が可能
- 間隔時系列に関する数理モデルを示した

この事実が何に依存するのか　　なぜ？は機会を改めたい

全体のまとめ

ビッグデータに内在する物理学?

言語に内在する数理的な不変性/普遍性

- 不変：時間安定な構造
- 普遍：異種サンプルに普遍な構造

エントロピー、冪則の二つの話題の最前線を示した

ご清聴ありがとうございました