令和２年度修士論文

単語の出現頻度に着目した
多言語テキストの構造解析

Structure analysis of multi-language texts focused on appearance frequency of words

大阪府立大学大学院
理学系研究科　物理科学専攻
非線形物理研究室
学籍番号2190302014
山本卓也

提出日：2021年2月26日

概要

頻度（サイズ）に関するべき乗則はZipf則と呼ばれ、名前の分布や地震の規模など様々な現象において確認されている。中でも文章中の単語の出現頻度のランクサイズ分布のZipf則は言語によらない文章構造の普遍的な側面と考えられている。しかし、Zipf則はあくまで分布則であるため、語順が考慮されておらず、文章としての特徴は結果には反映されない。そこで、本研究では単語の並びに着目した文章の数理的な解析を試みた。手法としては、文章内の単語列を単語ごとの出現頻度に応じたランク数列に変換し、そのリターンマップの分布に着目した。単語を無作為に再配置したデータ（サロゲートデータ）のリターンマップとの比較を行うことで、単語列の特徴を抽出した。解析対象には、英語・オランダ語・ドイツ語・フランス語・フィンランド語・ハンガリー語・イタリア語・ポルトガル語・スペイン語・日本語の10カ国語のテキストを各14冊ずつ計140冊を選出した。リターンマップの分布を特徴づける二種類のオーダーパラメータとして、サロゲートデータとの距離及び相関係数を定義し、テキストの比較を定量的に行った。その結果、（1）ほぼ全ての言語において隣接単語のランクに負の相関が見られることと、（2）同一言語テキスト同士のリターンマップに見られる特徴の類似性などが得られた。また、その結果を利用して言語ごとのリターンマップの特徴を再現するようなランク数列モデルの提案も行った。モデルには、べき乗則の成立過程としてよく知られるYule過程をベースに採用し、さらに上位ランクと下位ランクを分ける切替ランクを定義し、既出単語の再出現確率を隣接単語のランクに応じて重み付けを行った。重み付けの度合いも隣接ランクに応じて線形に変動するモデル（重み線形切替Yuleモデル）と度合いが一定であるモデル（重み二値切替Yuleモデル）の二つを提案した。モデルによって生成されたランク数列に対するリターンマップ解析を行い、オーダーパラメータ平面上で10言語のリターンマップと比較した結果、後者のモデルにおいて重み付けのパラメータを変化させることで言語ごとのHDMの特徴をある程度再現することができた。次に、英語テキスト75冊を一つの集団としてみなして2種類の解析（単語の非一様性と語数別のZipf則の検証）を行った結果として頻出単語と珍しい単語の切替過程が重要な要因となっていることが示唆された。最後にこれまでの解析結果を参考に人間の単語選択に対する考察として「二辞書モデル」を提案したい。本研究により、隣接単語間の相関に着目したことで言語テキストの特徴を抽出することができ、さらに頻出単語と珍しい単語で性質の違いが示唆された。

本論文に登場する用語・パラメーター一覧

用語・パラメータ	記号	説明
確率密度関数	p(x)	任意の要素 x と $x+dx$ の間のサイズを取る確率
累積分布関数	P(x)	任意の要素 x 以上のサイズを取る確率
べき指数	α	べき乗則の指数
総単語数	W	文章内に出現する総単語数
総語彙数	V	文章内に出現する単語の総語彙数
サイズ	s_i	単語 i の文章内での出現回数（頻度）
ランク	$r(s_i)$	サイズが s_i の単語 i に対応するランク
ランク列	r_n	文頭から数えて n 番目に登場した単語のランク
対数ランク列	R_n	対数を取ったランク列 r_n。$(R_n = log_{10} r_n)$
区間数	N_b	HDM の区間数
データ数	$f(R_i, R_j)$	HDM の区間 (i,j) 内のデータの個数
カラースケールの相対頻度	$B(i,j)$	データ数 $f(R_i, R_j)$ を総単語数 W で割った値
縦列ごとの KL 情報量	D_i	HDM の縦列 i における $B(i,j)$ と $\tilde{B}(i,j)$ の分布の差異
テキストの新単語発生確率	$P_{nw}(n)$	未登場の単語が発生した確率、n 番時の語彙数 V_n を単語数 W_n で割った値
相関係数	C	HDM 上でのデータ分布の偏り
相対確率密度	$\psi(i,j)$	カラースケール $B(i,j)$ と $\tilde{B}(i,j)$ に対する相対確率密度
ユークリッド距離	D	カラースケール $B(i,j)$ と $\tilde{B}(i,j)$ のユークリッド距離
モデルの新単語発生確率	γ	既出でない全く新しい単語が発生する確率
総語彙数	V_{Max}	文章内に出現する単語の種類数
初期単語数	V_s	Yule 過程の開始までに既に一度ずつ発生していた単語
乱数の種	S	プログラムの乱数発生に用いた整数
切替ランク	r_c	頻出単語とそうでない単語を区別するためのランク
確率変動パラメータ	λ	重み線形切替 Yule モデルで切替ランクに応じて拡張 Yule 過程の挙動を変える変数
確率変動パラメータ	Δ	重み二値切替 Yule モデルで切替ランクに応じて拡張 Yule 過程の挙動を変える変数
既存単語発生確率	$P_{n,j}$	n 番目に単語 j が出現する確率
集団全体の総単語数	W_{all}	テキスト集団全体に登場する総単語数
集団全体の総語彙数	V_{all}	テキスト集団全体に登場する単語の総語彙数
単語ごとの KL 情報量	$D_w(i)$	テキスト集団内での単語 i の出現分布の差異
テキスト別サイズ	$s_{i,j}$	テキスト j における単語 i の出現回数
集団内サイズ	S_i	テキスト集団内での単語 i の出現回数

目 次

1 序論

昔から多くの研究者たちがべき乗則という法則に対して関心を寄せてきた。べき乗則に従う現象としては名前の分布、論文の引用、ウェブ上での検索ヒット数、本の売り上げ、地震の規模、太陽フレアの強度、戦争の規模など広い分野に渡って確認されている [1]。最もよく知られているのは Zipf 則であり、それは文章内に出現する単語の出現頻度に関するものであり、その成立は文章の総単語数や総語彙数、言語によらず成立が報告されている。単語の出現頻度に関する他の性質としては単語のバースト性や Heaps 則などがあり、それらに対する研究も数多くなされている。

初めに、Zipf 則とは単語の出現頻度や都市の人口、名前の発生などのそれぞれの要素の大きさ（サイズ）x を両対数グラフにプロットしていった時、線形な形を取る、つまりべき乗則に従うという経験則である [2]。つまり、任意の要素が x と $x + dx$ の間の大きさをとることを表す確率密度関数 $p(x)$ を定義した場合

$$\ln p(x) = -\alpha \ln x + c \tag{1}$$

という関係が得られる。ここで、α と c は定数である。(1) 式の両辺を真数に戻すことで

$$p\left(x\right) = Cx^{-\alpha} \tag{2}$$

となり、べき乗則に従うことが示された。($C = \exp(c)$)

さらに、任意の要素の大きさが x 以上の形を取る確率を示す累積分布関数 $P(x)$ は

$$P(x) = \int_x^\infty p(x')dx' \tag{3}$$

で示されるので、(2) 式を代入すれば

$$P(x) = C\int_x^\infty x'^{-\alpha}dx' = \frac{C}{\alpha - 1}x^{-(\alpha-1)} \tag{4}$$

となり、$P(x)$ もまたべき乗則となる。先行研究から、べき指数 α は、

$$\alpha = 1 + n[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}}]^{-1} \tag{5}$$

で求められるとされている。n は要素の数を示す。しかし、式 (5) については様々な議論がなされており、本研究ではそれらの議論を基に α の導出を修正した。

べき乗則の発生過程はしばしば Yule 過程で説明される。Yule 過程とは、1. 多数存在する集合はそれぞれの要素に比例した割合で新しい要素を増やす、2. 定期的に新しい集合を発生させることである。Hayakawa らは日本人の名前の実データを解析し、そのべき乗則を再現するために Yule 過程を拡張したモデルを提案した [3]。結果としては Yule 過程を採用したことで実データによく似たべき乗則の分布が見られた。名前と単語では発生過程に大きな違いがあると考えられるが、べき乗則を再現するという点では Yule 過程は非常に有益であると言える。Sunehag[4] や Madsen ら [5] は TF-IDF やディリクレ多項分布を利用した単語生成モデルの提案を行ったがどちらも単語のバースト性を再現するためのものであった。

これまではべき乗則ばかりに着目してきたが、これはあくまで分布則であるため、人間の精神活動の産物として文章の意味を成立させる重要な単語の語順や文法などは全く考慮されていない。それに着目した数値的な解析も勿論存在し、様々な手法を用いて数値的な解析が行われている。例えば、Schenkel らは Peng らによる DNA 配列に関する先行研究 [6] を基に、文章内の文字列を解析し、分布則以外の特徴を研究した [7]。手法としては、アルファベット（26 文字）といくつかの記号を、5 ビットの 2 進数（$2^5 = 32 > 26$）で表すことにより、文章を 0 と 1 のみの一次元ランダムウォークモデルとみなし、パワースペクトルを調べた。その中では、小説テキストのみならずコンピュータープログラムや聖書なども解析されており、聖書の文字列が強い相関を持つことなどが主張されている。

このように様々な研究者 [8][9][10] が、異なる手法で文章の長距離相関に注目してきた、特に Altmann は単語や文字だけではなく、トピックや母音・子音を単語や文字と同じように解析した。このように長距離相関に比べて、単語間のより近い相関についての理解は十分になされているとは言えない。本研究では、隣接する単語の並びに着目する形で文章構造を解析する方法を提案し、それを通して多言語テキストの数理的な特徴づけを目的とする。我々は Zipf 則の普遍的な成立な性質を調べた後で、文字のデータを数理的データにとして扱うために、Zipf 則にも用いられたランクを単語に割り当てることで、文章内の単語列を数列とみなし、数列に対する時系列解析を様々な手法を用いて行った。また、そこから得られた結果とべき乗則の双方を再現するために、隣接単語のランクの大きさに応じて拡張された Yule 過程の挙動を変更するランク数列発生モデルを提案したい。結果として、単語は上位・下位ランクを切替えるランクの存在が隣接単語の相関には重要であることがわかった。切替ランクの存在を検証するために Kamada らの日本人の姓名の分布の都道府県に対する非一様性に関する先行研究 [13] を参考に、英語テキスト集団内における各テキストへの単語の非一様性を調べる。さらに英語テキスト集団を利用した解析としてランダムに抽出した単語群に対して語数別のランクサイズを比較したことで出現頻度に依存した単語の性質の違いが示唆された。最後にこれまでの解析結果を参考に人間の単語選択に対する考察として「二辞書モデル」を提案したい。

本論文の構成を示す。2 章では英語テキストにおける Zipf 則の成立を検証し、式 (5) によるフィッティングと修正したべき指数のフィッティングを比較した。3 章ではリターンマップ解析により言語テキストの隣接単語間における相関を様々な手法で調べた。4 章で言語テキストのリターンマップを再現するようなランク数列の提案を行い、5 章で英語テキスト集団に対する解析と「二辞書モデル」の提案を行った。

2　英語テキストのZipf則とべき指数の検証

2.1　Zipf則の検証

統計量として文章内に出現する単語の総数は総単語数W、単語の種類数は総語彙数V、単語iの出現回数はサイズs_i、サイズs_iに対するランクとして、s_i回以上出現する単語の種類数を用いる。このとき、総単語数W、総語彙数Vのテキスト、出現する単語のサイズs、ランク$r(s)$との間には以下の関係が成立する。

文章における単語の出現回数の最小値は通常1であるから、

$$V = r(s_{min}) = r(1) \tag{6}$$

$$W = \sum s \tag{7}$$

Zipf則において、単語を要素としてみなすので式(1)で用いられたxに対して

$$x = s \tag{8}$$

$$r(s) = P(s)V \tag{9}$$

の関係が成立し、ランク$r(s)$とサイズsとの間には

$$r(s) = Vs^{-(\alpha-1)} \tag{10}$$

というべき乗則の形が表される。

以降、実際の英語テキスト15冊分テキストデータをプロットし、Zipf則の成立を検証する。次にべき指数αを式(5)を修正したものから決定し、最後にプロットとフィッティングした結果を示す。15冊の英語テキストには、"Project Gutenberg"から主に小説テキストを中心に抽出したものを使用した。テキストには、以下の前処理を行った。

1. アルファベットの大文字（ABCD...）は全て小文字（abcd...）に変換
2. コンマなどの記号は全て削除　ex)．,"?!＆＃。、など
3. 数字(0123456789)は削除

これらの処理を施した後で、各テキストの単語のサイズとランクを決定した。各テキストの総単語数、総語彙数、最頻出単語（ランク1位の単語）とそのサイズを表2にまとめた。

総単語数、総語彙数ともに大きく違うテキストであるが、最頻出単語が「the」であることがわかった。英語テキストにおいて上位にくる単語はtheやaといった冠詞、of・in・toなどの前置詞、it・he・thatのような代名詞、is・areなどの動詞が見られた。さらに、下位には人や物の名前などの名詞や副詞、動詞の活用形が多く見られた。一例としてMobyDickのサイズランク分布の一部を表3に示す。前述の通り、ランク最下位は総語彙数に対応している。実際に、このデータを両対数プロットしたものが図1である。横軸がサイズ、縦軸がランクに対応している。グラフの線型の形から、ランクがサイズのべき乗に従っている、すなわち、Zipf則が総単語や総語彙数に関係なく成立することが見てとれる。

テキスト名	総単語数 W	総語彙数 V	最頻出単語 (サイズ)
MobyDick	207007	19403	the (13920)
Pride and Prejudice	122141	6312	the (4331)
Dracula	161776	9321	the (7881)
A tale of two cities	137135	9786	the (8024)
Base-Ball	39004	3518	the (3335)
Brothers of Karamazov	351750	121184	the (15173)
Frankenstein	75059	71110	the (4187)
Grimm's Fairy Tales	101179	4768	the (6980)
Gulliver's travels	102124	8178	the (5777)
The Adventures of Sherlock Holmes	104319	8333	the (5601)
Adventures of Huckleberry	111934	6162	the (6350)
Iliad	172544	11310	the (14218)
The Rupublic	217720	10265	the (15404)
The Romance of lust	189957	8263	the (7944)
The adventures of Tomsawyer	71283	7325	the (3711)

表 1: 15 冊の英語テキストデータの諸元。

2.2 べき指数 α の決定

Zipf 則におけるべき指数 α は式 (5) から得られ、$x_{min} = 1$ とすると英語版 MobyDick のべき指数 $\alpha = 2.26$ となる。しかし、実際にフィッティング（図 2）してみると、サイズが大きくなるにつれて実データとの差が開いてしまっていることが確認できる。先行研究によると、これは (5) 式の $x_{min} = 1$ を修正する必要がある。そこで、x_{min} の値を 1 から 10 まで順に大きくしていき、それぞれの値に対するべき指数 α を求めた（図 3）。これより、x_{min} を大きくするにつれて、べき指数 α はある値で収束するような振る舞いを示した。この結果から隣接するべき指数 α 同士の差が初めて ± 0.015 以内に収まった時の平均値を新しいべき指数 α とした。MobyDick であれば $x_{min} = 4$ と $x_{min} = 5$ でのべき指数 α の値を平均した。新たに求めたべき指数は $\alpha = 2.03$ となった。同じ方法で求めた他の英語テキストのべき指数 α を表 4 にまとめた。Zipf 則におけるべき指数 α は 2 前後の値を取ることがわかる。こちらも本研究で使用した全てのテキストに対して総単語数、総語彙数、言語によらず同様の傾向を確認できた (付録 A.6)。修正したべき指数 $\alpha = 2.01$ を用いて、MobyDick の実データとのフィッティングを行った。図 4 より、よくフィッティングされていることが確認できる。これより、Zipf 則の成立が確認され、

$$r(s) = Vs^{-(\alpha-1)} = 19403s^{-1.01} \tag{11}$$

が示された。他の 14 冊の英語テキストに対しても、2.1.2 で決定したべき指数 α を用いてフィッティングを行った。結果として、MobyDick 同様によくフィッティングされていることが確認された。また、これからの解析に使用する英語以外の言語に対しても同様の結果が得られた (付録 A.6 参照)。

ランク r_s	サイズ s	単語
1	13920	the
2	6382	of
3	6228	and
4	4516	a
5	4484	to
⋮	⋮	⋮
10	1921	i
11	1757	but
12	1724	he
⋮	⋮	⋮
9497	2	glasshopper
9497	2	mathmatics
⋮	⋮	⋮
19403	1	eggs
19403	1	sypathecally
19403	1	damped
⋮	⋮	⋮

表 2: 英語版 MobyDick のサイズランク分布。

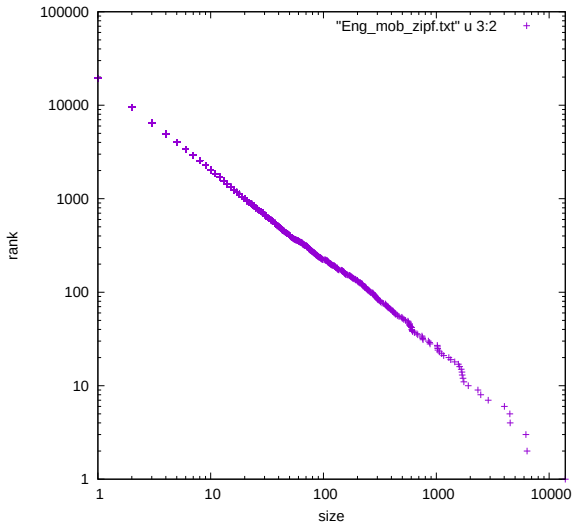

図 1: 英語版 MobyDick データの両対数プロット。横軸が単語のサイズ、縦軸がそれに応じたランク。

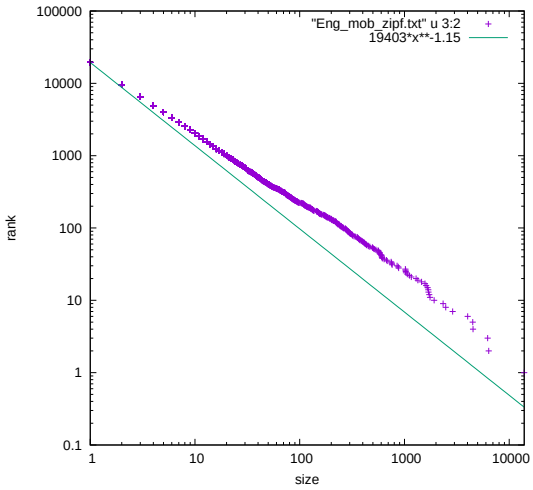

図 2: べき指数 $\alpha = 2.25$ でのフィッティング。横軸が単語のサイズ、縦軸がそれに応じたランク。

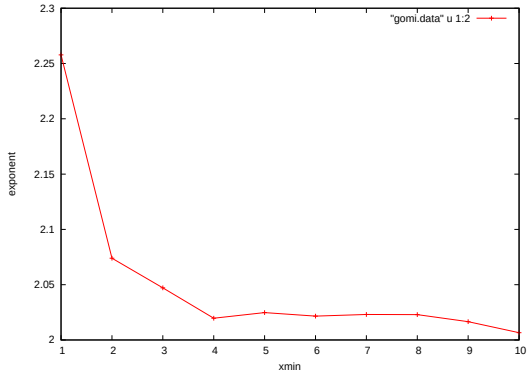

図 3: $x_{min} = 1 \sim 10$ でのべき指数 α の値。

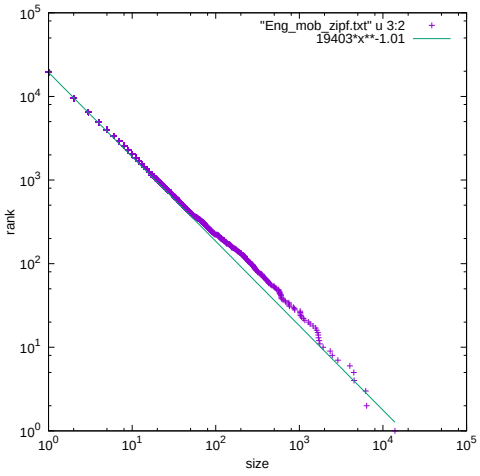

図 4: MobyDick の修正したべき指数 $\alpha = 2.01$ によるフィッティング。横軸が単語のサイズ、縦軸がそれに応じたランク。

テキスト名	元の α	修正した α
MobyDick	2.26	2.03
Pride and Prejudice	1.83	1.84
Dracula	2.00	1.84
A tale of two cities	2.02	1.95
Base-Ball	2.10	1.89
Brothers of Karamazov	1.76	1.80
Frankenstein	2.05	2.06
Grimm's Fairy Tales	1.79	1.81
Gulliver's travels	2.02	2.00
The Adventures of Sherlock Holmes	2.10	1.96
Adventures of Huckleberry	1.92	1.84
Iliad	1.87	1.92
The Rupublic	1.87	1.88
The Romance of lust	1.79	1.82
The adventures of Tomsawyer	2.19	2.01

表 3: 英語テキストのべき指数 α。横軸が最小値サイズ x_{min}、縦軸がべき指数 α。

3　言語テキストに対するリターンマップ解析

3.1　英語テキストのリターンマップ

Zipf 則は分布則であり、単語の出現の順番は考慮されていない。そこで本研究では単語の並びに着目した。解析対象には、"Project Gutenberg"から再び 14 冊の英語テキストを小説を中心に抽出した（Zipf 則の検証とテキストの重複あり）。単語の順番に関する傾向を定量化する文章解析を行うために、まずテキストデータを数列的なデータに変換しなければならない。本研究では、文章を構成する各単語を Zipf 則で用いた出現頻度のランクに変換した。英語版 MobyDick を例に用いるならば、the は 1 に、of は 2 に、grasshopper や mathmatics などの名詞は 9497 に変換される。文章を先頭から数えて n 番目の単語のランク r_n とする。これより、文章内の単語列がランク数列 r_n に変換された。語彙数は最上位ランクに、総単語数は末項に対応する。ランクの定義により、同じサイズの単語には同じランクが割り当てられる。その意味で単語とランクは一対一対応ではないが、これによりランク列 r_n を解析した結果は文章内の単語の並びを解析した結果に対応するとみなした。ランク列 r_n を単純にプロットしてみたが周期性のようなものはみられなかった（図 5）ので、ランク列全体を解析するのではなく、最も相関が強いと考えられる隣接するランク列同士、つまり r_n と r_{n+1} の関係を調べることにした。

隣接項間の相関を視覚化する手段としてリターンマップ、すなわち x 軸に r_n をとり、y 軸に r_{n+1} をプロットしたマップを用いた。例えば、of the の順番に単語が並んでいる時であれば $(2, 1)$ の座標をプロットする。図 6（左）の英語版 MobyDick（総単語数 207007、総語彙数 19403）のランク列のリターンマップから、ランク数列 r_n の取りうる値が離散的なことと、そのプロットが非常に偏った分布であることが見て取れる。さらに、重複回数もこのマップでは表されていない。これらの問題を解決するために、ランクの対数 $R_n = log_{10} r_n$ に対して、二次元ヒストグラムをとった。ヒストグラムの区間の数 $N_b = 10$ で統一し、カラースケールは区間内のデータの個数に対応している。リターンマップの左下の領域はランク上位（頻出単語）が連続していることを示し、左上の領域はランクの上位の次にランク下位（珍しい単語）が来ることを意味している。図 6（中）を見ると左上にやや濃い程度の偏りが見られる。ここで我々は、ランダムに再配置したデータ（サロゲートデータ）との差をとることで、同じランクサイズを持つが文章としての意味を持たないランダムな単語列に見えなかった文章の単語の並びの特徴を抽出することに成功した。図 6（右）がサロゲートデータのマップであり、マップに偏りはほとんど見られない。以降、文章データとサロゲートデータの相対頻度の差を取ったリターンマップを Histogram Difference return Map（HDM）と呼ぶ。

図 7 は MobyDick（英語版）の HDM であり、カラースケールの相対頻度は元データに対して

$$B(i,j) = \frac{\text{区間内のデータの個数 } f(R_i, R_j)}{\text{総単語数 } W} \quad i,j = 1,2,3...,10 \tag{12}$$

で定義し、サロゲートデータのカラースケールの相対頻度 $\tilde{B}(i,j)$ との差をとった。具体的には j 番目の区間には $\frac{j}{N_b}\log_{10} V \leq R_n < \frac{j+1}{N_b}\log_{10} V$ を満たす単語が入る。（対数をとる前のランク r_n で表すと $V^{\frac{j}{N_b}} \leq r_n < V^{\frac{j+1}{N_b}}$ となる。）カラースケールが赤い区間は元データの相対頻度 $B(i,j)$ がサロゲートデータの相対頻度 $\tilde{B}(i,j)$ より大きいことを表

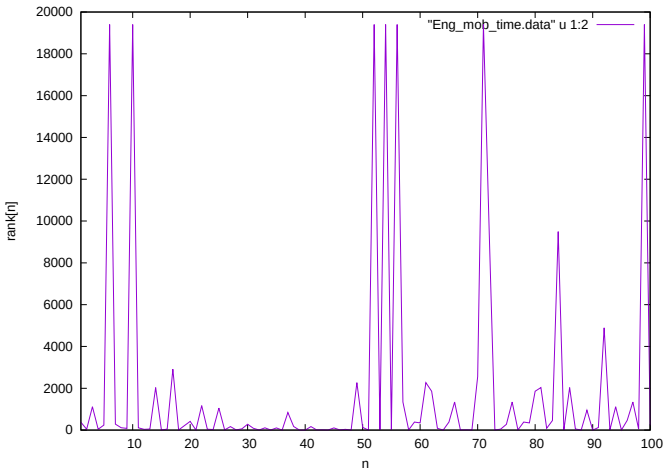

図 5: 英語版 MobyDick のランク列プロット。縦軸がランク数列 r_n、横軸が項番号 n。第 100 項までに周期性のようなものは見られなかった。

す。青い区間はその逆を表す。これより、赤い区間の分布が元データにおける隣接項間の特徴を抽出しているとした。例えば、図 7 の $(i,j)=(2,1)$ は薄い赤であるがその部分は 3 ～ 5 位の単語の次に 1 ～ 2 位の単語が並んだ場合を数えている。単語のランクは上位から the,and,a,of,to,in,that... であるので、この場合 to the や of the のような「前置詞+the」の組み合わせが元データとして頻繁に出現していることが分かる。逆に、$(i,j)=(1,2)$ のような組み合わせはほとんど見られないので青くなっている。

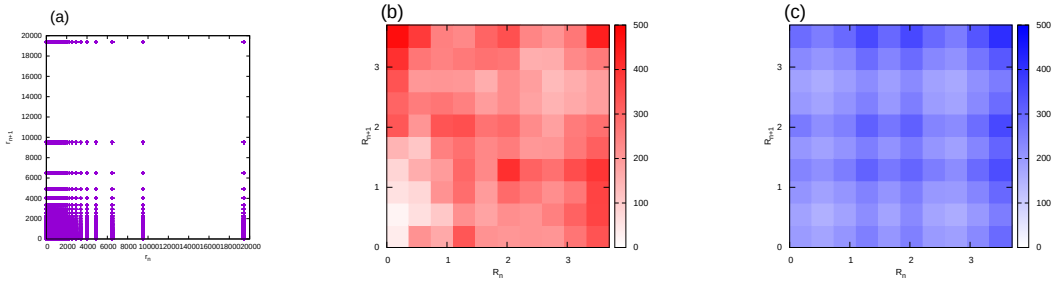

図 6: 英語版 MobyDick のランク列のリターンマップ。(a) 一般的なリターンマップ、両対数ヒストグラムリターンマップ（底は 10）(b) 元データ、(c) サロゲートデータ。横軸が対数ランク列 R_n、縦軸が隣接項の対数ランク列 R_{n+1}。カラーバーは区間内の頻度数。

英語テキストの HDM を比較する。テキストは"Project Gutenberg"から 14 冊を選出したため、フリーテキストの総語彙数や総単語数に統一性はない。これらに対して、条件を統一した後に HDM を作成した。条件としては各テキストは先頭から語彙数が 5000 になるまでを使用した。総語彙数がランクの最下位に対応しているため、ランク数列 r_n に割り当てられる値は 1 ～ 5000 までとなる。さらに、ヒストグラムの区間の数 N_b を縦横ともに 10 で統一したことにより、カラースケールの値 $B(i,j)$ を同程度にすることができ、定量的な比較が容易になった。図 8 に英語テキスト 1 4 冊分の HDM を載せた。図 8 左上は 14 冊分の英語テキストの HDM の平均をとったものである。そのカラースケールは $\langle B(i,j)\rangle = \frac{1}{k}B_k(i,j), k=1,2,3...14$ である。どのマップにおいても対角領域（$R_n = R_{n+1}$

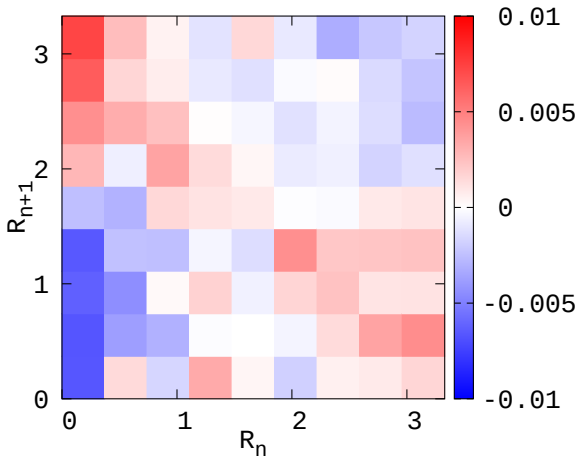

図 7: MobyDick の HDM（底は 10、総語彙数は 5000）。横軸が対数ランク列 R_n、縦軸が隣接項の対数ランク列 R_{n+1}。カラーバーは区間内の相対頻度。

に近い部分）ではサロゲートデータの相対頻度の方が大きく、左隅と非対角成分（対角領域でない領域）では元データの相対頻度の方が大きくなった。これから、英語テキストでは、頻出単語の次に頻出単語、頻出単語の次に珍しい単語、珍しい単語の次に頻出単語の並びが多く存在するという結果が得られた。

英語版 MobyDick の HDM の区間の数を変更した（図 9）。区間の数が小さいと細かい構造の違いが見えにくく、大きくなるにつれて区間の中にプロットされないことが増えてきて、白い区間が増える傾向にある。

3.2　6 言語テキストのリターンマップ

ここまで英語テキストの単語の並びに関する構造解析の結果を得た。これより、これは英語テキストのみに共通する構造なのだろうか？それとも、他の言語で書かれたテキストも共通の構造を持つのだろうか？共通の構造が存在するならば、どのような構造をもつのだろうか？という問題がある。これを調べるためにも、リターンマップ解析を行った。なぜなら、テキストがランク数列 r_n に変換されたことで同じ条件、方法で異なる言語の比較が可能だからである。これより、多言語テキストのリターンマップを比較し、それぞれの言語における構造を調べた。

使用する言語として、英語・オランダ語・ドイツ語・フランス語・フィンランド語・日本語の 6ヶ国語を選んだ。日本語以外のテキストは英語同様"Project Gutenberg"から主に小説のフリーテキストを使用した。日本語テキストは"青空文庫"から同様に小説を中心にフリーテキストを使用した。また、英語テキストと同様に条件を統一したテキストを解析した。条件は英語と同様に各テキストは先頭から語彙数が 5000 になるまでを使用し、ヒストグラムの区間の数 N_b を縦横ともに 10 で統一した。中には、Adventures of Tom Sawyer（英語）と Die Abenteuer Tom Sawyers（ドイツ語）のように同一作品の異なる言語で翻訳されたテキストも含まれる。また、日本語文章は空白によって単語が分かれていないた

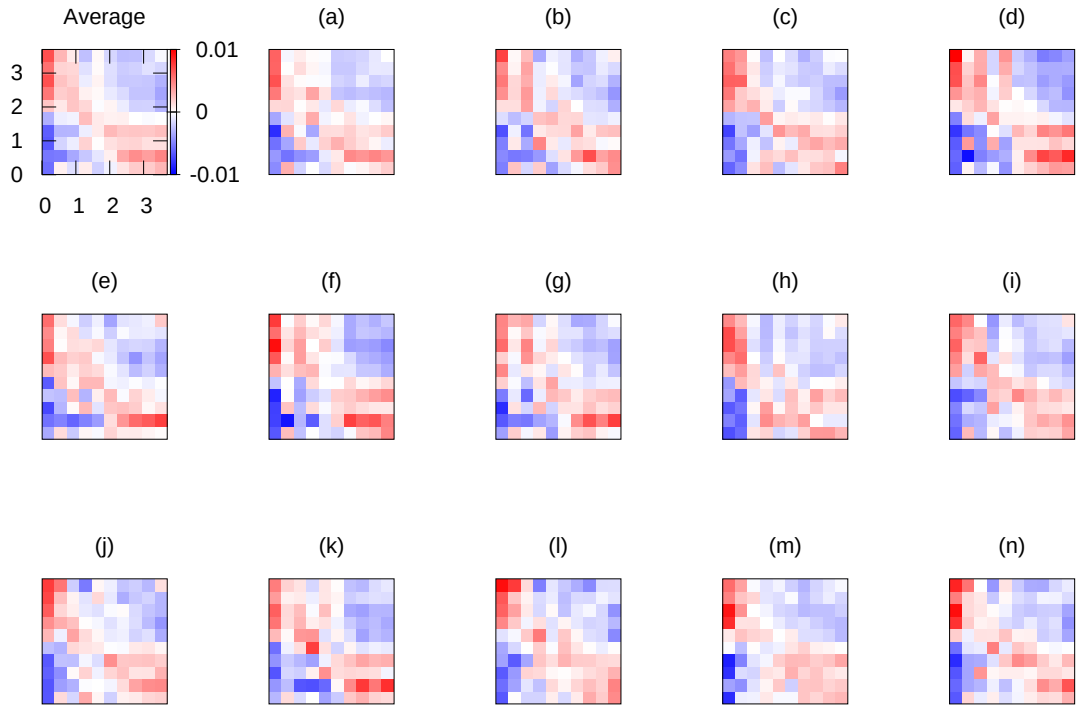

図 8: 英語の全 14 冊分の HDM。(a)The hound of the Baskervilles, (b)Brothers of Kramazov, (C)Crime and Punishment, (d)Frankenstein, (e)The picture of Dorian Gray, (f)Gulliver's travels, (g)The house of the dead, (h)Adventures of Huckleberry, (i)War and Peace, (j)MobyDick, (k)Prime and prejudice, (l)The adventures of Tomsawyears, (m)Treasure Island, (n)The tale of two cities.

め形態素解析ソフト"mecab"により、分かち書き（品詞ごとに分解）を行った。「吾輩は猫である」という文章も例に用いれば、「吾輩（名詞）」「は（助詞）」「猫（名詞）」「で（助詞）」「ある（動詞）」のように分解される。

各言語 14 冊ずつ計 84 冊分のテキストデータを解析した結果、非対角領域に元データの分布が多いなどの傾向は見られるが、その度合いは必ずしも同じ程度ではなかった。しかし、それぞれ言語ごとで HDM を比べると非常によく似た構造が見られるものもあった（図 23）。例えば、フランス語（図 12）は英語と同様に非対角領域に元データが大きく偏った分布を持ち、対してフィンランド語（図 13）は偏りもサロゲートデータとの差も小さい。オランダ語（図 10）とドイツ語（図 11）は非対角領域に元データの分布が多く見られるが、全体的にサロゲートデータとの差は大きくない、すなわち単語の並びに偏りがあまり見られない。日本語（図 14）はフランス語以上に元データが大きな偏りを持ち、さらに頻出単語同士の並びは全く見られない。この結果と英語テキストの結果から、言語はそれぞれの言語ごとに特有の共通な構造を持つことが示唆される。

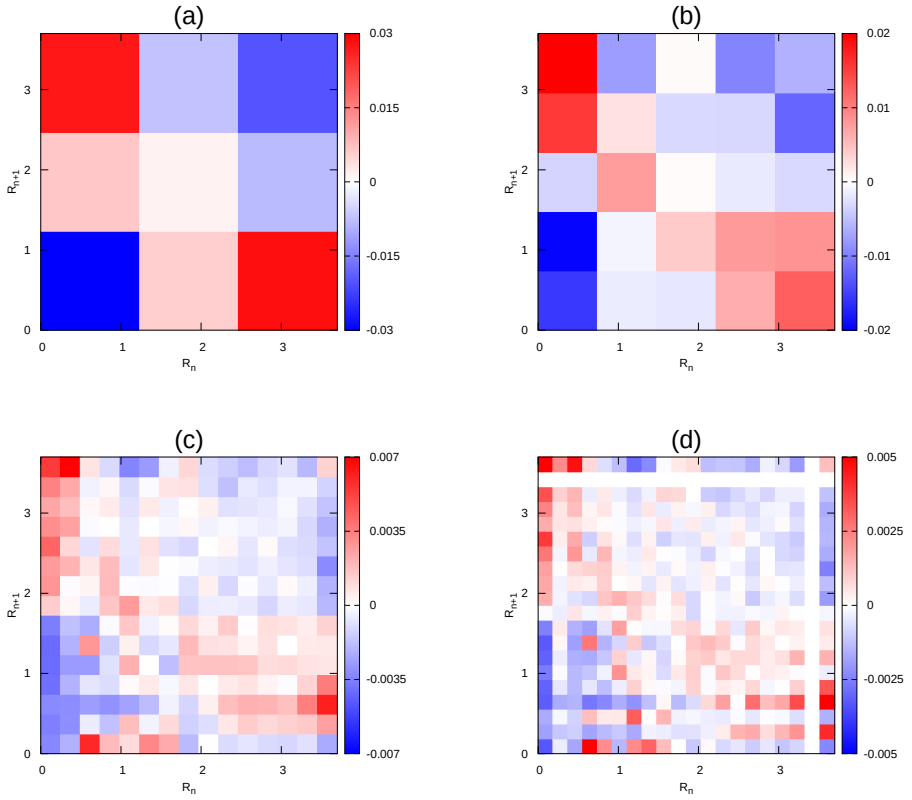

図 9: 区間の異なる英語版 MobyDick の HDM。(a)3×3、(b)5×5、(c)15×15、(d)20×20。横軸が対数ランク列 R_n、縦軸が隣接項の対数ランク列 R_{n+1}。カラーバーは区間内の相対頻度。

3.3 言語リターンマップの特徴づけ

3.3.1 言語 HDM の階層的クラスタ解析

前節の結果から、言語 HDM は言語ごとに特有な共通の構造の存在が示唆された。これを定量的に調べるための手法として、初めに階層的クラスター解析を行った。階層的クラスタ解析とは、値が距離が近いデータからデンドログラム（樹形図）で順番にまとめていく（クラスタリング）分析方法のことである。本研究では、HDM の区間の値 $B(i,j)-\tilde{B}(i,j)$ をベクトルの成分とみなした $100(=10\times10)$ の成分を持つベクトル間の距離が近いテキストからクラスタリングしていった。距離にはユークリッド距離を、距離測定法には Ward 法を採用した。84 冊分のテキスト（6 言語 14 冊）に対するクラスタ解析の結果から、オランダ語・ドイツ語・英語で分岐があるものも 6 ヶ国語全てで言語ごとのクラスタができていることがわかる (図 16)。

3.3.2 Kullback-Leibler 情報量を用いた切換ランクの推定

次に言語全体に共通する HDM の特徴が非対角領域への分布の偏りであることから市松模様のような四分割のブロック構造を、その偏り具合が言語ごとに異なることからブロッ

ク構造の境界線が言語ごとに共通であることを仮定した。そして、境界線となる対数ランク列 $R_n(= log_{10} r_n)$ を頻出単語（ランク上位の単語）と珍しい単語（下位単語）を区別するような対数切換ランク R_c として定義し、言語ごとの切換ランク r_c を調べることとした。

対数切換ランク R_c の推定には元データのカラーバーの値 $B(i,j)$ とサロゲートデータのカラーバーの値 $\tilde{B}(i,j)$ が小さくなるところを求めたい。そのために Kullback-Leibler 情報量を用いた。Kullback-Leibler 情報量とはそれぞれの確率分布が異なる対象の差を表すものであり、$B(i,j)$ と $\tilde{B}(i,j)$ を確率分布のようにみなして HDM の縦列 i ごとの Kullback-Leibler 情報量 D_i

$$D_i = \sum_{j=1}^{10} B(i,j) \cdot \ln\{B(i,j)/\tilde{B}(i,j)\} \tag{13}$$

を求めた。これにより、KL 情報量 D_i の値が減少していくにつれては元データ（赤い分布）とサロゲートデータ（青い分布）の差も小さくなっていくと考えられる。

6 言語ごとに KL 情報量をプロットしていくと (図 17)、どの言語でも第 1 列から減少し、途中で再び増加する傾向が見られる。我々はこの傾向から最小値をとる列の対数ランク列 R_n が対数切換ランク R_c ではないかと考えた。これを r_n に戻すことで得たそれぞれの切換ランク r_c は英語・オランダ語・ドイツ語で 30 位前後、フランス語で 71 位前後、フィンランド語で 388 位前後、日本語で 12 位前後である。図 23 を見ると、この結果は適当であるように見えるが定量的な評価はできなかった。

3.3.3 新単語発生確率の線形性

後述するランク数列生成モデルのために HDM 以外での特徴づけを行なった。そこで注目したのが未登場の単語が発生した確率、新単語発生確率 $P_{nw}(n)$ であり、n 番目の単語が発生する時の総語彙数 V_n と同時の総単語数 W_n を用いて $P_{nw}(n) = \frac{V_n}{W_n}$ と定義した。図 18 の縦軸に $P_{nw}(n)$ を、横軸に単語順（ランク列項数）n をプロットした。HDM に合わせて n の対数をとっているため片対数グラフになっている。これより、言語全体に共通の傾向として新単語発生確率 $P_{nw}(n)$ は一度減少しはじめると線形に減少するという傾向が見られる。そして、その減少具合（線形の勾配）は言語ごとに等しいように考えられる。これから、単語は最初期に異なるものがいくつか出現し、一度単語の再出現が見られると他の単語も同様に再出現しはじめ、その確率は言語ごとに一定であると考察した。

3.3.4 オーダーパラメータ D と C の導入

これまで様々な方法で HDM の特徴づけを行なったが定量的な評価が満足になされているとは言えない。ここでもう一度 HDM のデータの偏り具合や差を定量的に特徴づけるために、HDM に対してオーダーパラメータを二種類定義した。一つ目は、相関係数 C である。まず元データとサロゲートデータのカラースケール $B(i,j)$,$\tilde{B}(i,j)$ に対して相対確率密度

$$\psi(i,j) = \left(\sum_{m,n} \frac{B(m,n)}{\tilde{B}(m,n)} \right)^{-1} \cdot \frac{B(i,j)}{\tilde{B}(i,j))} \tag{14}$$

を定義すると、元データの区間ごとの平均 $\langle R_i \rangle$ は

$$\langle R_i \rangle = \sum_{m,n} R_m \cdot \psi(m,n) \tag{15}$$

で表される。よって、HDM のデータの偏り具合である相関係数 C は以下のように定義できる。

$$C \equiv \frac{\sum_{i,j}(R_i - \langle R_i \rangle)(R_j - \langle R_j \rangle)\psi(i,j)}{\sqrt{\sum_{i,j}(R_i - \langle R_i \rangle)^2 \psi(i,j)}\sqrt{\sum_{i,j}(R_j - \langle R_j \rangle)^2 \psi(i,j)}} \tag{16}$$

これは HDM の（赤い）分布が対角線近くに分布している場合には正の値を、非対角領域に多く分布している場合には負の値をとる。

また、二つ目のオーダーパラメータとして HDM における文章データとサロゲートデータ間のユークリッド距離 D は

$$D \equiv \sqrt{\sum_{i,j}(B(i,j) - \tilde{B}(i,j))^2} \tag{17}$$

となり、HDM の色（赤青共に）濃いほど大きい値をとる。

3.4 DC 平面上での 10 言語 HDM の比較

ここで 4 言語テキスト（ポルトガル語、イタリア語、スペイン語・ハンガリー語）をそれぞれ 14 冊ずつ追加した（図 19-22）。狙いとしては我々の語順に着目した解析と形態的言語類型論との関係を調べるためである。形態的類型論とは言語を孤立語・膠着語・屈折語の 3 つの類型に分類するもので、先ほどまで使用していた 6 言語においてはフィンランド語と日本語が膠着語に該当する。膠着語とは自立語（名詞や動詞）に機能語（助詞や接辞）がくっついて文になる言語のことである。しかし、HDM の構造において変わった構造が見られたのはフィンランド語だけである。ここで我々は日本語は mecab による「分かち書き」を行なったことで膠着語としての性質が見えにくくなったのではないかと考えた。そこで膠着語であるハンガリー語とそれ以外の類型に分類される 3 言語を追加して、ハンガリー語とフィンランド語の類似性やその他の言語の特徴を調べた。

図 24 に全 140 冊分の相関係数 C と距離 D を示した。ほとんど全ての文章の相関係数 C の値が負の領域に分布しているので、HDM の分布が非対角領域にある、つまり頻出単語と珍しい単語が交互に来る傾向が 10 言語テキストには共通の構造であることが定量的に示された。またドイツ語とフィンランド語とハンガリー語は元データとサロゲートデータとの差が大きくないことやフランス語・イタリア語・ポルトガル語・スペイン語・日本語が大きな偏りを持つこと、言語ごとに特有の特徴を持つことなども DC 平面から見て取れる。

英語版 MobyDick を例に用いて解析位置を文頭以外に変更した場合や隣接項以外の相関を調べた場合の結果についてもオーダーパラメータ D と C を用いた (付録 A.1 A.2 参照)。また、4 言語（ポルトガル語、イタリア語、スペイン語・ハンガリー語）テキストを追加する前に行った特徴づけ（3.3.1 節〜3.3.3 節）の結果は付録 A.3〜A.5 に示す。

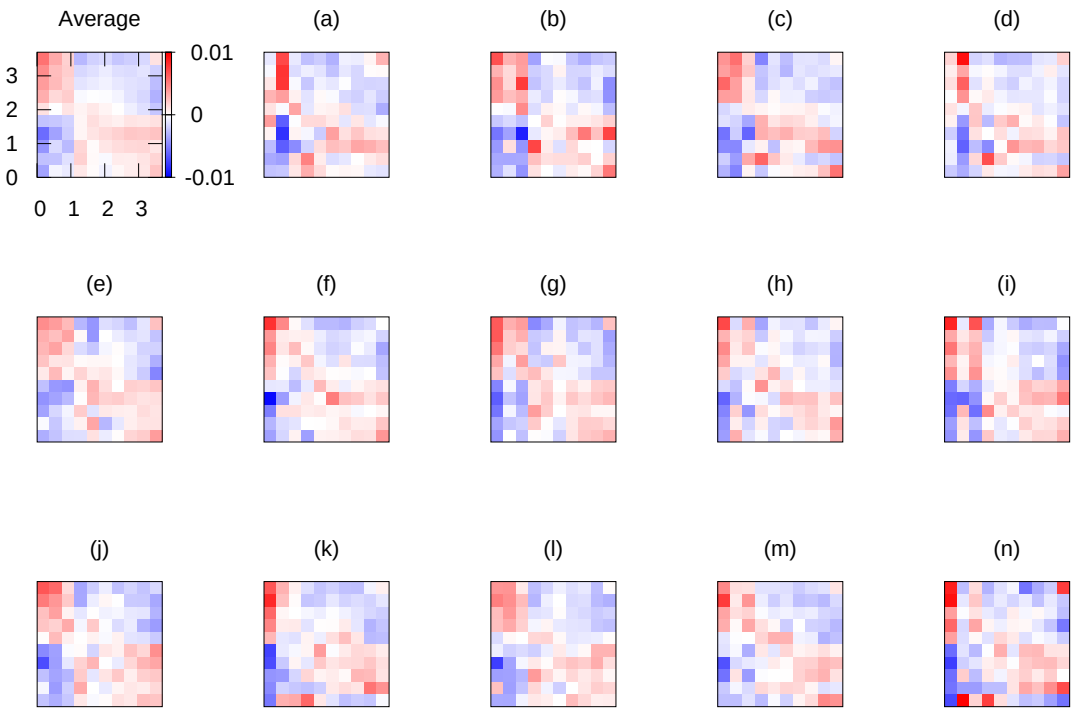

図 10: オランダ語の全 14 冊分の HDM。(a)Het portret van Dorian Gray, (b)De verrezen Gulliver, (c)De Lotgevallen van Tom Sawyer, (d)Beatrice, (e)Eline Vere Een Haagsche roman, (f)Gosta Berling, (g)De legende en de heldhaftige, vroolijke en roemrijke daden van Uilenspiegel en Lamme Goedzak in Vlaanderenland en elders, (h)Per luchtschip de Argonaut naar Mars, (i)De wijzen van het Oosten Brahmanisme, Boeddhisme, Chineesche philosophie, Mazdeisme, (j)Martelaren van Rusland, (k)Het boek van Siman den Javaan Een roman van rijst, dividend en menschelijkheid, (l)De Talisman of Richard Leeuwenhard in Palestina, (m)Vonken, (n)Beowulf Angelsaksisch volksepos vertaald in stafrijm en met inleiding en aanteekeningen voorzien.

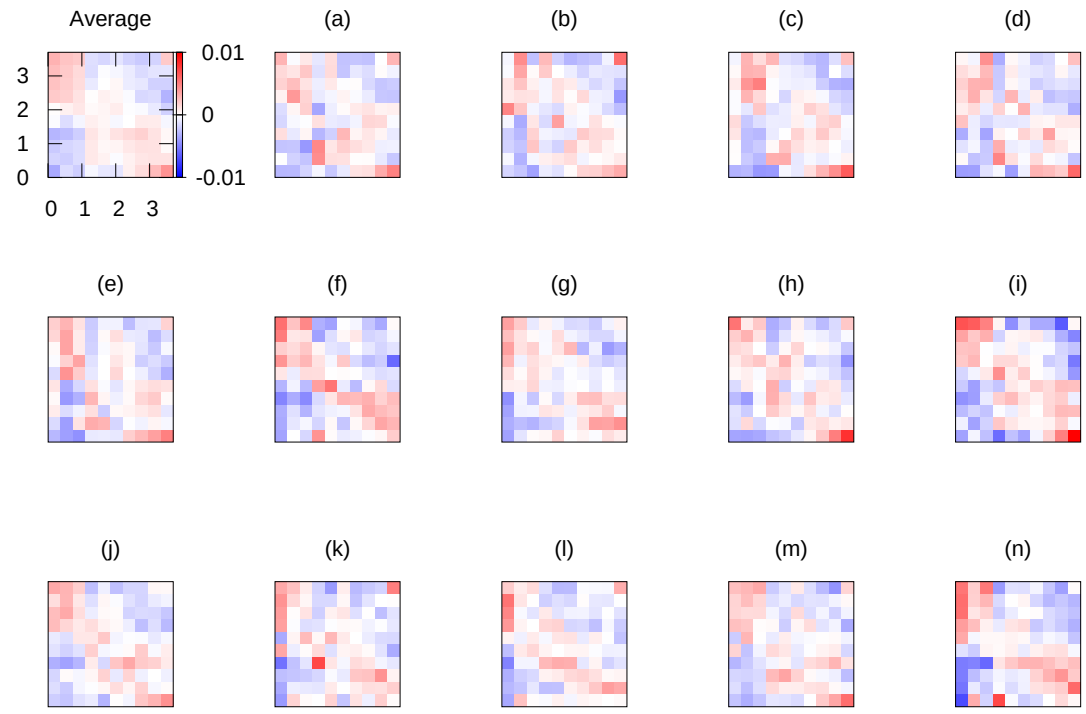

図 11: ドイツ語の全 14 冊分の HDM。(a)Der Weihnachtsabend Eine Geistergeschichte, (b)Das Bildnis des Dorian Gray, (c)Die Abenteuer Tom Sawyers, (d)Die Schatzinsel: Roman, (e)Unsichtbare Bande Erzahlungen, (f)Quer Durch Borneo Ergebnisse seiner Reisen in den Jahren 1894, 1896-97 und 1898-1900; Erster Teil, (g)Die Abtissin von Castro, (h)Drei Monate Fabrikarbeiter und Handwerksbursche Eine praktische Studie, (i)Geschichte von England seit der Thronbesteigung Jakob's des Zweiten Zweiter Band, (j)Lichtenstein, (k)Die Weltensegler. Drei Jahre auf dem Mars. 1910(l)Die Religion innerhalb der Grenzen der blo β en Vernunft Text der Ausgabe 1793, mit Beifugung der Abweichungen der Ausgabe 1794, (m)Die Klerisei, (n)Der Weltkrieg, II. Band Vom Kriegsausbruch bis zum uneingeschrankten U-Bootkrieg.

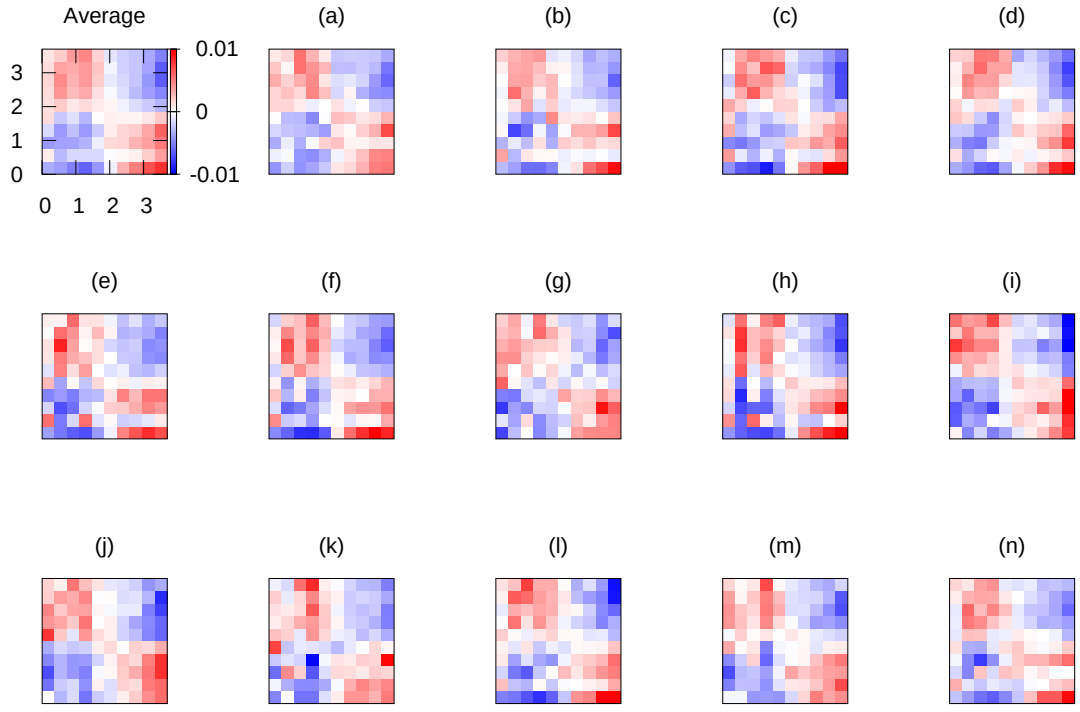

図 12: フランス語の全 14 冊分の HDM。(a)Cantique de Noel, (b)Le portrait de Dorian Gray, (c)Les voyages de Gulliver, (d)Souvenirs de la maison des morts, (e)Le livre de la Jungle, (f)Vie de Christophe Colomb, (g)Psychologie de l'education, (h)Thais, (i)La Force Le Temps et la Vie, (j)Le Dragon Imperial, (k)Les mille et une nuits Tome premier, (l)Quentin Durward, (m)Souvenirs d'egotisme autobiographie et lettres inedites publiees par Casimir Stryienski, (n)La case de l'oncle Tom ou vie des negres en Amerique.

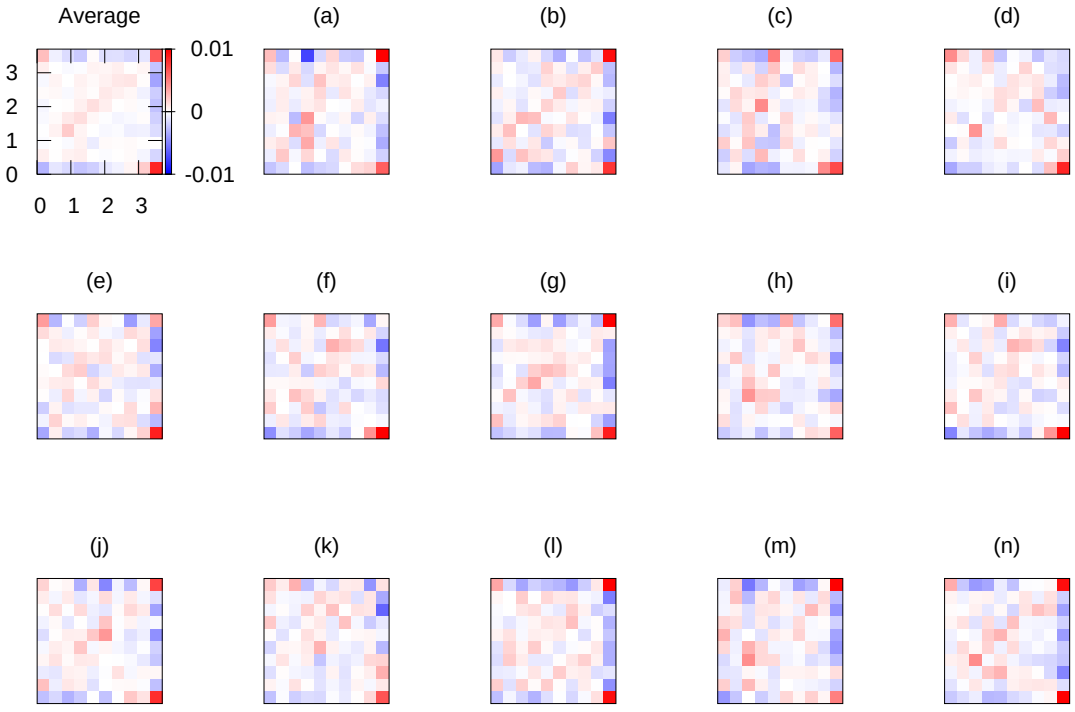

図 13: フィンランド語の全 14 冊分の HDM。(a)Baskervillen koira, (b)Rikos ja rangaistus, (c)Gorgias, (d)Gulliverin matkat kaukaisilla mailla, (e)Muistelmia kuolleesta talosta, (f)Huckleberry Finnin (Tom Sawyerin toverin) seikkailut, (g)Viidakkopoika, (h)Ylpeys ja ennakkoluulo, (i)Tom Sawyersin seikkailut, (j)Aarresaari, (k)Gil Blas Santillanalaisen elamanvaiheet, (l)Jean-Christophe X Uusi tyopaiva, (m)Kun nukkuja heraa Romaani, (n)Vtanha tarina Montrosesta.

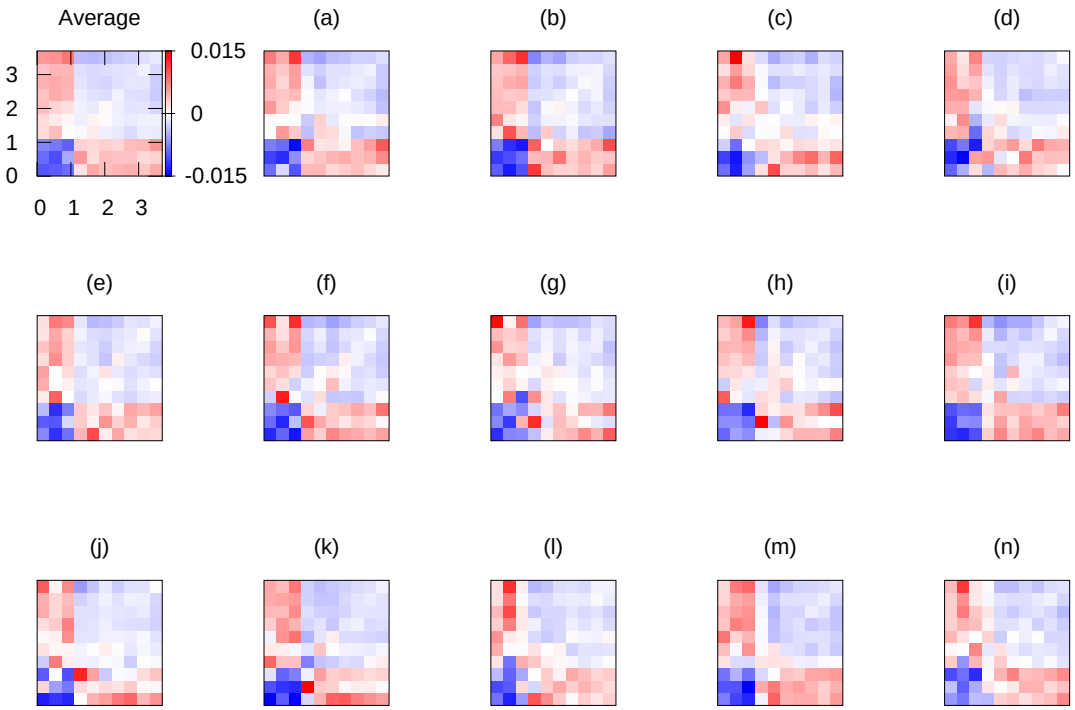

図 14: 日本語の全 14 冊分の HDM。(a) カラマーゾフの兄弟、(b) クリスマスキャロルの夜、(c) フランケンシュタイン、(d) ガリバー旅行記、(e) 宝島、(f) 二都物語、(g) 人間失格、(h) 痴人の愛、(i) 夜明け前　第一部上、(j) 地名の研究、(k) 樋口一葉訳　源氏物語「若菜」、(l) こころ、(m) 三四郎、(n) 坊ちゃん。

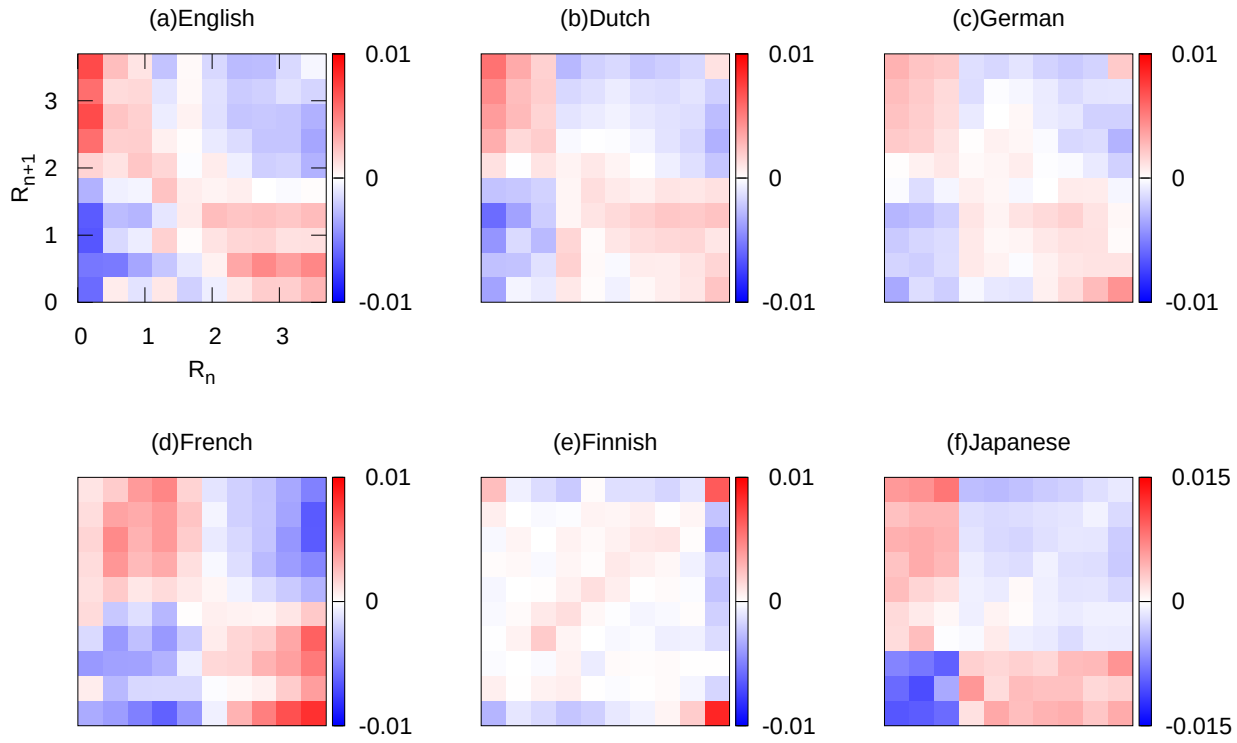

図 15: 6 カ国語テキストをそれぞれ 14 冊分を平均化した HDM。(a) 英語、(b) オランダ語、(c) ドイツ語、(d) フランス語、(e) フィンランド語、(f) 日本語。

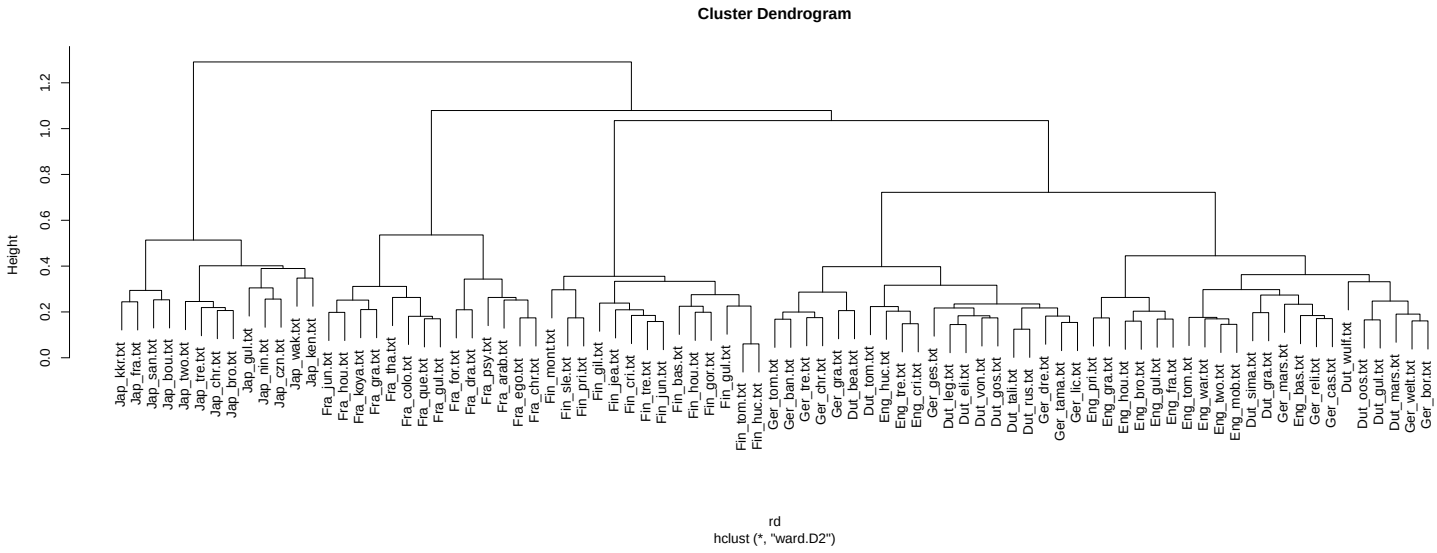

図 16: Ward 法でクラスタリングした 6 言語テキストのデンドログラム。縦軸がクラスタ間距離。

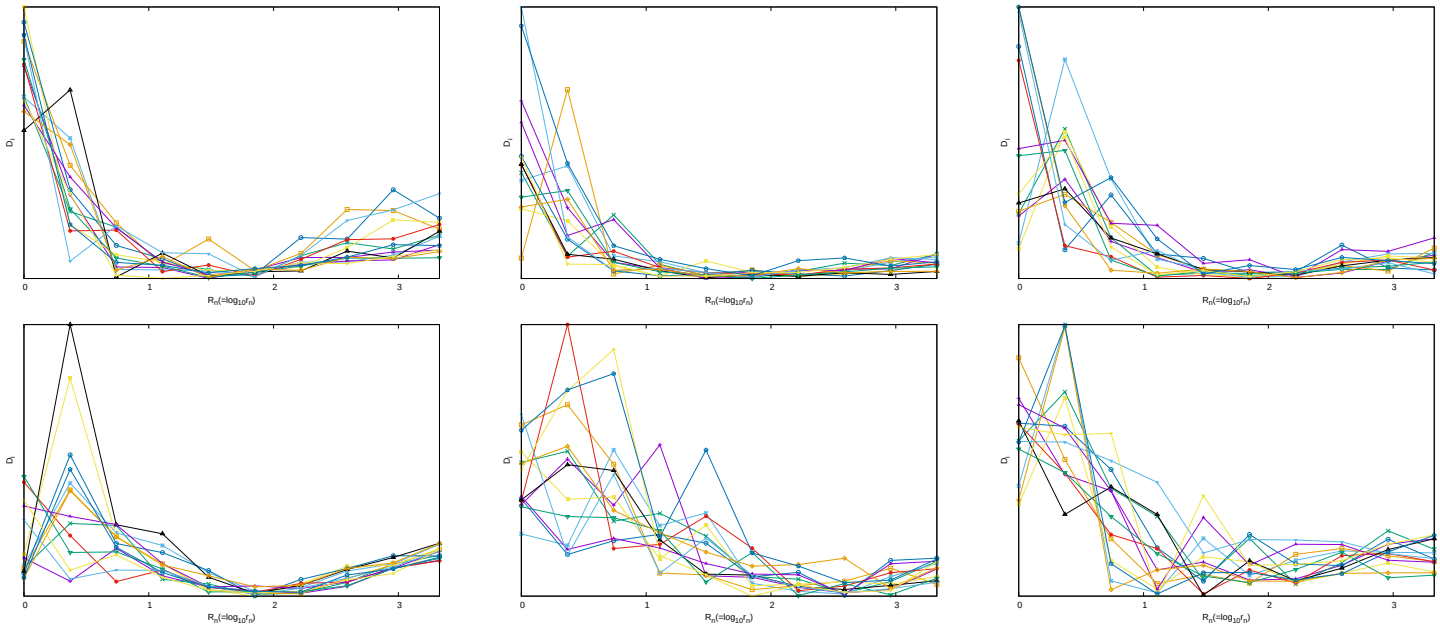

図 17: 6 言語の KL 情報量。上列左から英語、オランダ語、ドイツ語、下列左からフランス語、フィンランド語、日本語。横軸が対数ランク列 R_n、縦軸が HDM の縦列ごとの KL 情報量 D_i。

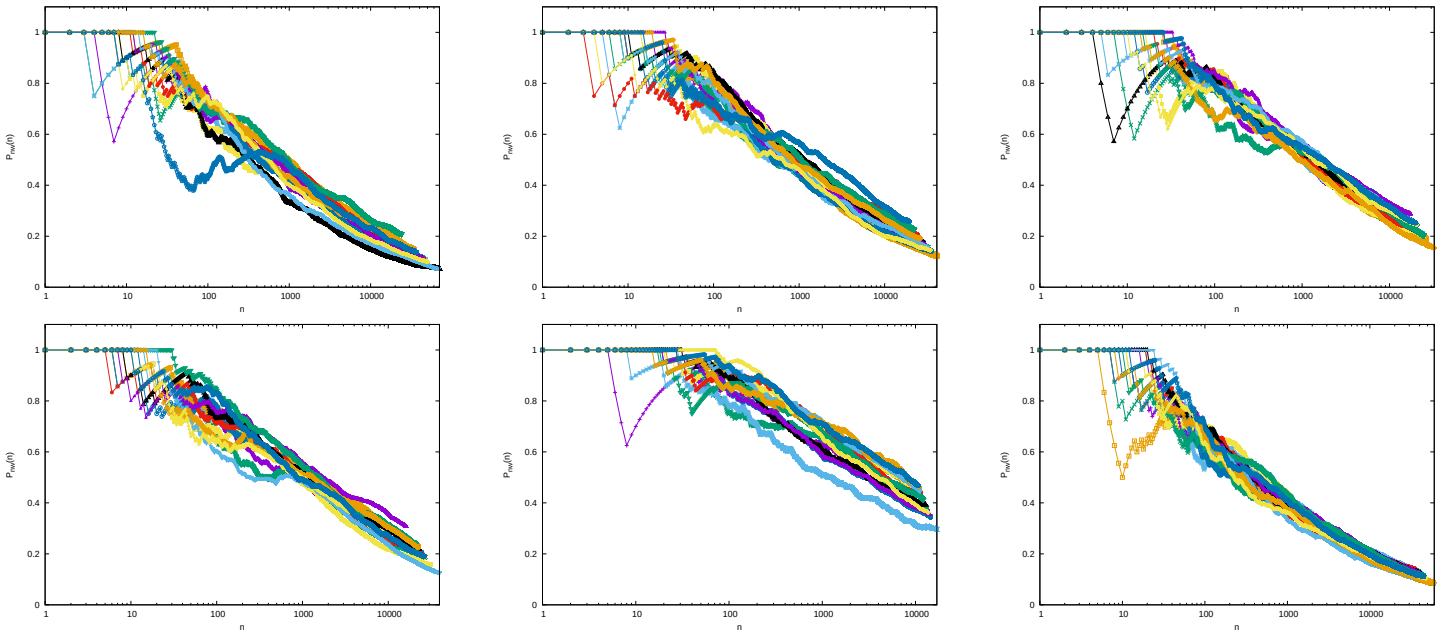

図 18: 6 言語の新単語発生確率 $P_{nw}(n)$。上列左から英語、オランダ語、ドイツ語、下列左からフランス語、フィンランド語、日本語。横軸が項番号 n、縦軸が新単語発生確率 $P_{nw}(n)$。

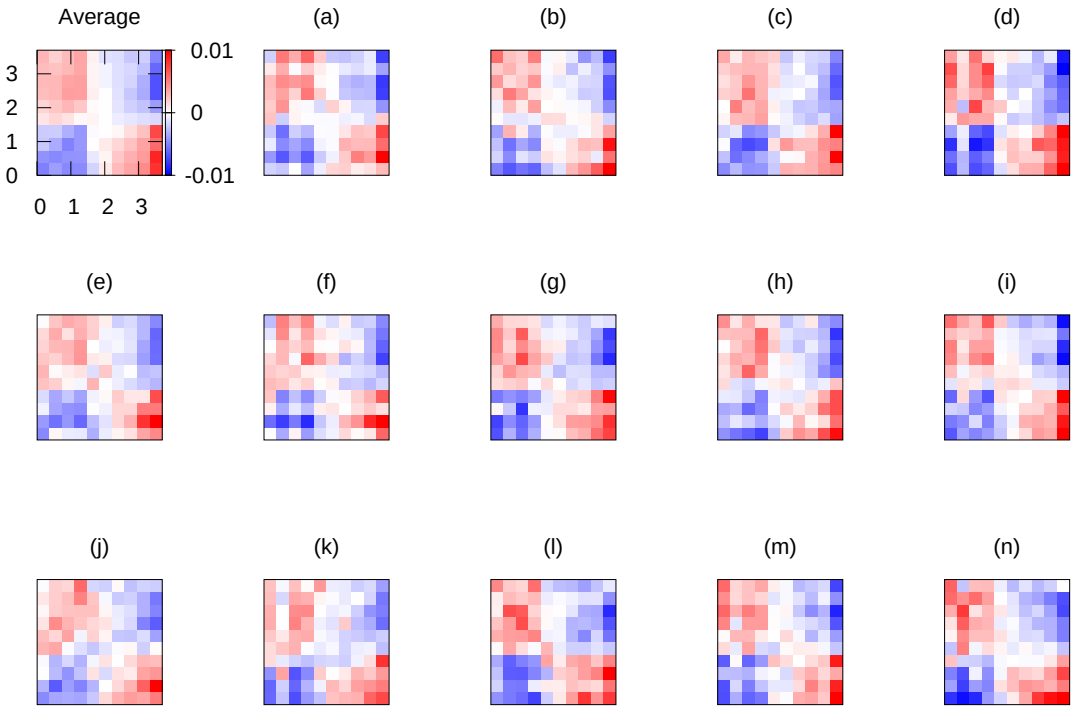

図 19: ポルトガル語の全 14 冊分の HDM。(a)Ambicoes: Romance, (b)Amor Crioulo vida argentina, (c)Portugal e Brazil: emigracao e colonisacao, (d)Os deputados brasileiros nas Cortes Geraes de 1821, (e)Os fidalgos da Casa Mourisca Chronica da aldeia, (f)Os Filhos do Padre Anselmo, (g)Historia de Portugal: Tomo I, (h)Viagem ao norte do Brazil feita nos annos 1613 a 1614, pelo Padre Ivo D'Evreux, (i)A Morgadinha dos Cannaviaes (Chronica da aldeia), (j)Obras Completas de Luis de Camoes, Tomo II, (k)A Reforma, (l)Resumo elementar de archeologia christa,(m)Os Trabalhadores do Mar, (n)Tres capitaes.

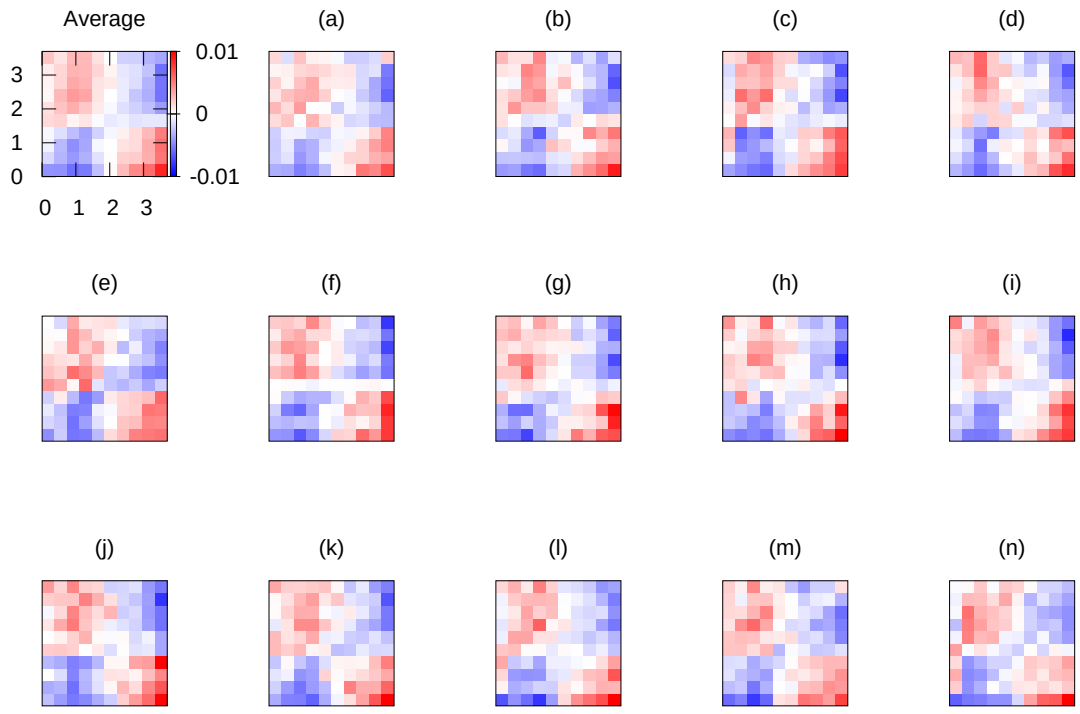

図 20: イタリア語の全 14 冊分の HDM。(a)Fra Tommaso Campanella, Vol. 1 la sua congiura, i suoi processi e la sua pazzia, (b)Della storia d'Italia dalle origini fino ai nostri giorni, sommario. v. 1, (c)Ettore Fieramosca: ossia, La disfida di Barletta, (d)In faccia al destino, (e)Fra Tommaso Campanella, Vol. 2 la sua congiura, i suoi processi e la sua pazzia Language, (f)La guerra del Vespro Siciliano vol. 1 Un periodo delle storie Siciliane del secolo XIII, (g)La guerra del Vespro Siciliano vol. 2 Un periodo delle storie Siciliane del secolo XIII, (h)Lettere di Lodovico Ariosto Con prefazione storico-critica, documenti e note, (i)I mesi dell'anno ebraico, (j)Niccolo de' Lapi; ovvero, i Palleschi e i Piagnonia, (k)Le rive della Bormida nel 1794, (l)Novelle, (m)La scienza in cucina e l'arte di mangiar bene Manuale pratico per le famiglie, (n)Orlando Furioso.

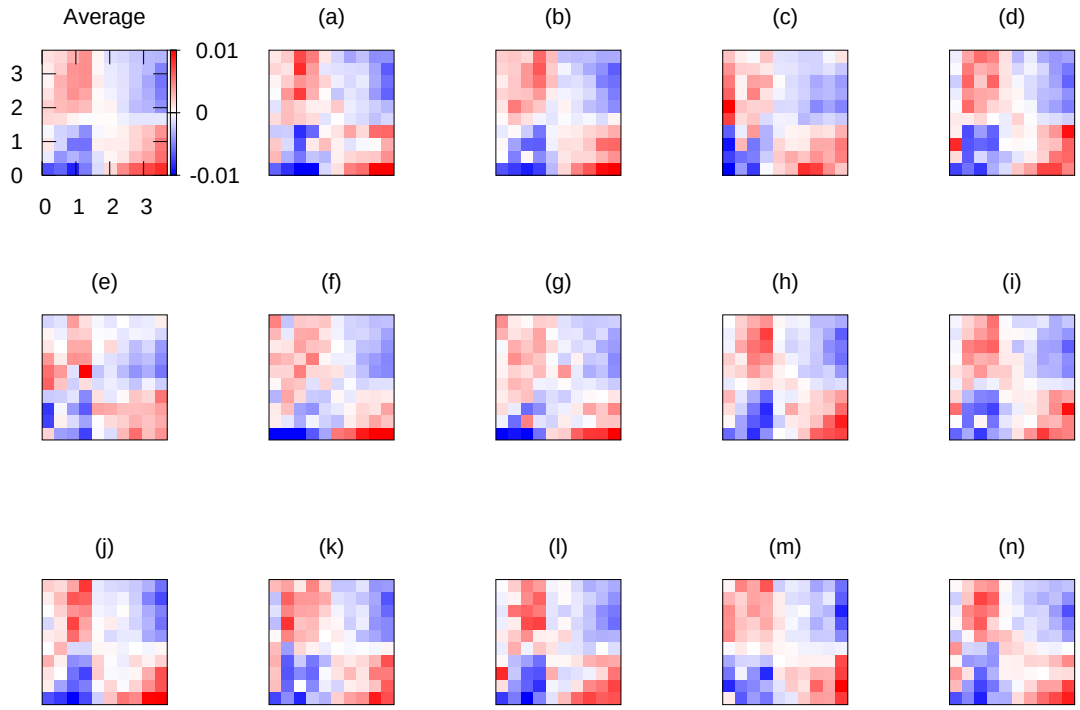

図 21: スペイン語の全 14 冊分の HDM。(a)El amor, el dandysmo y la intriga, (b)La Argentina La conquista del Rio de La Plata. Poema historico, (c)Cocina moderna, (d)El Criterio, (e)Filosofia Fundamental, (f)Historia de Venezuela, Tomo I, (g)Historia de Venezuela, Tomo II, (h)Las noches mejicanas, (i)Los Merodeadores de Fronteras, (j)Orlando Furioso, Tomo I, (k)Un paseo por Paris, retratos al natural, (l)El Protestantismo comparado con el Catolicismo en sus relaciones con la Civilizacion Europea (Vols 1-2), (m)La Regenta, (n)Su unico hijo.

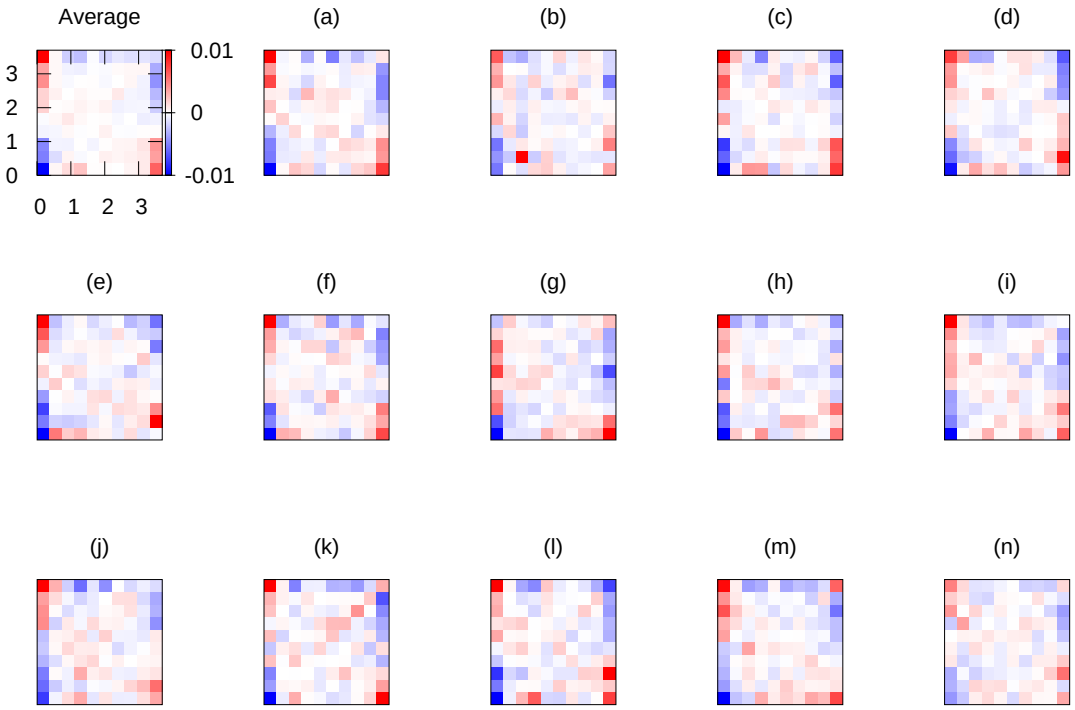

図 22: ハンガリー語の全 14 冊分の HDM。(a)Alomvilag: Elbeszelesek, (b)Elbeszelesek, (c)A lathatatlan ember: Regeny, (d)Szazadunk magyar irodalma kepekben: Szechenyi follepesetol a kiegyezesig, (e)Vezeto elmek: Irodalmi karcolatok, (f)Edes anyafoldem! : Egy nep s egy ember tortenete (1. kotet), (g)Az erkolcsi vilag, (h)Midas kiraly, (i)Torpek es oriasok, (j)A voros regina: regeny, (k)Edes anyafoldem! : Egy nep s egy ember tortenete (2. kotet), (l)Furcsa emberek: Elbeszelesek, (m)Nepmesek Heves- es Jasz-Nagykun-Szolnok-megyebol; Magyar nepkoltesi gyujtemeny 9. kotet, (n)Vegzetes tevedes: Regeny.

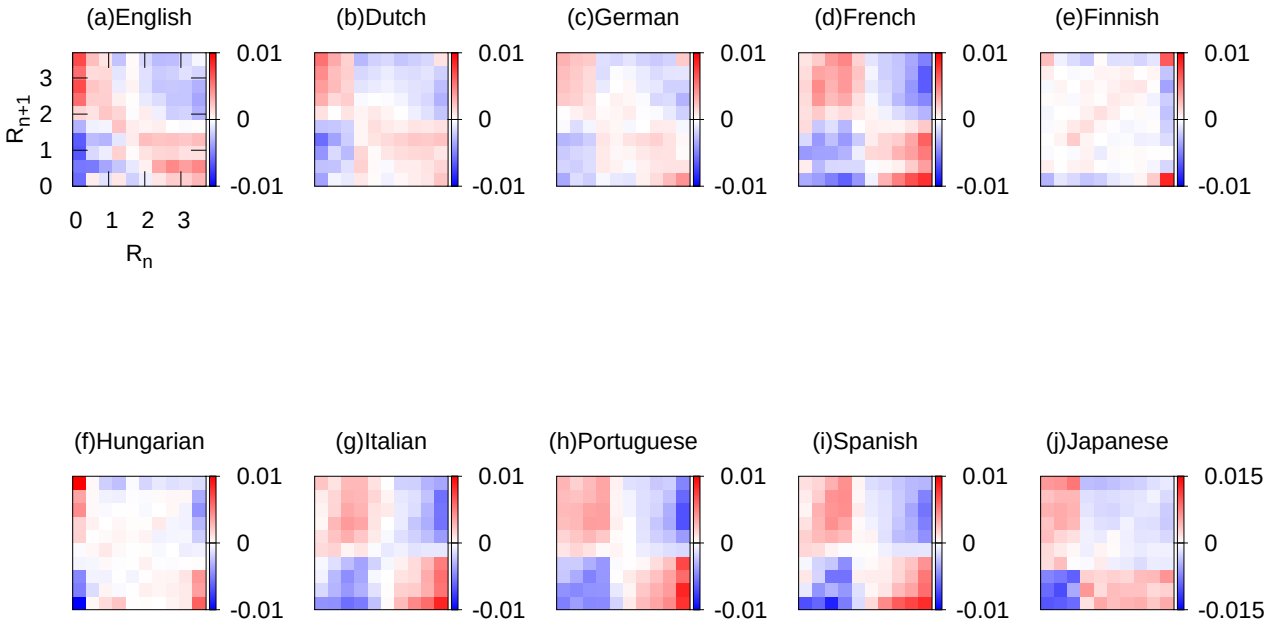

図 23: 10 カ国語テキストをそれぞれ 14 冊分を平均化した HDM。(a) 英語、(b) オランダ語、(c) ドイツ語、(d) フランス語、(e) フィンランド語、(f) イタリア語、(h) ポルトガル語、(i) スペイン語、(j) 日本語。

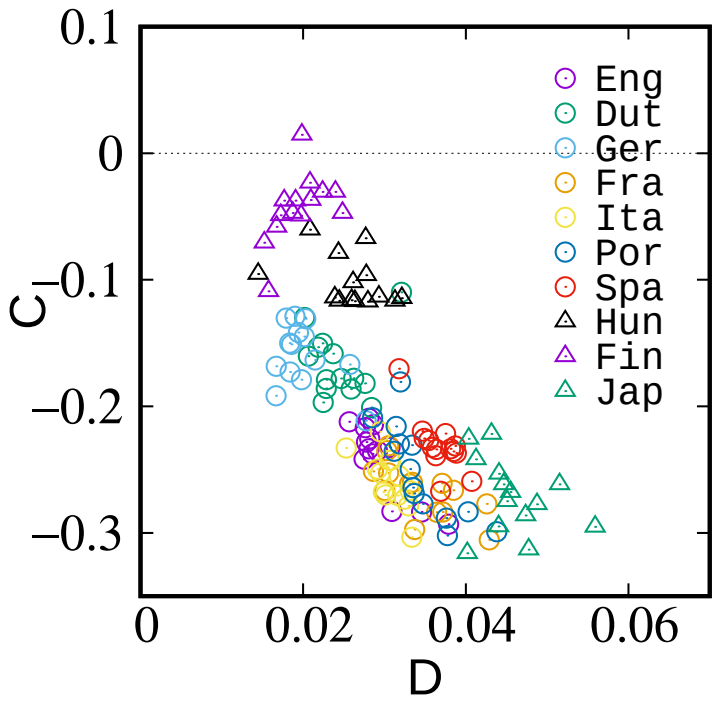

図 24: 10 言語 14 冊分の DC 平面。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。

4　ランク数列生成モデルの提案と検証

リターンマップ解析によって得られた言語ごとの特徴を持ったランク数列モデルを提案し、DC 平面上で言語 HDM との比較を行なった。

4.1　拡張 Yule 過程

ランク数列モデルの提案に際して、べき乗則の成立は不可欠である。それを考慮し、単語の発生プロセスには Yule 過程を採用した。

Yule 過程の 2 つの過程は以下のように定義される。

- 一定の割合 γ で新種が発生する。
- とある種 j が再び発生する確率はそのサイズ s_j に比例する。

また、HDM 上に見られる非対角領域の相関の再現を行うために、1. 頻出単語（ランク上位）に隣接する単語のランクはランダム、2. 珍しい単語（ランク下位）に隣接する単語は上位ランクになりやすいという 2 つの特徴を Yule 過程から生成されるランク数列に与えたい。そこで Hayakawa らによる Yule 過程の拡張 [3] を参考に、n 番目に単語 j が再び出現する確率 $P_{n,j}$ は $n-1$ 番目に出現した単語 j のサイズ $s_{n-1,j(k)}$ をそれぞれパラメータ β 乗することにより、単語の再出現過程に重み付けをした（式 (18)）。$\beta > 1$ であれば、より大きいサイズ $s_{n-1,j}$ を持つ単語への重み付けの影響が強くなり、それらがよく選ばれるようになる。$\beta = 1$ であれば、単純な Yule 過程となり、それぞれの単語のサイズ S_j に比例した割合で選ばれる。$\beta < 1$ であれば、$\beta = 0$ に近づくにつれて、重み付けの影響が弱くなることで出現のランダム性が増していき、$\beta = 0$ で完全なランダムで単語 j が選ばれる。隣接単語のランクに応じて β の値を変える二つのモデルの提案とその検証を 4.2~4.3 節で行う。

$$P_{n,j} = \frac{s_{n-1,j}^{\beta}}{\sum_k s_{n-1,k}^{\beta}} \tag{18}$$

4.2　重み線形切替 Yule モデル

モデルのパラメータを以下のように定義した。

パラメータ	記号	説明
新単語発生確率	γ	既出でない全く新しい単語が発生する確率
総語彙数	V_{Max}	文章内に出現する単語の種類数
初期単語数	V_s	Yule 過程の開始までに既に一度ずつ発生していた単語
乱数の種	S	プログラムの乱数発生に用いた整数
切替ランク	r_c	頻出単語とそうでない単語を区別するためのランク
確率変動パラメータ	λ	切替ランクに応じて拡張 Yule 過程の振る舞いを変更する変数

表 4: モデルのパラメータ。

β の変化量は確率変動パラメータ λ により線形に変化させた。以下でそれについてまとめた。

- 語数が V_s になるまで、異なる単語が 1 語ずつ出現。
- 語数が V_s に達してからは拡張 Yule 過程に従って、単語が出現。
  1. 一定の確率 γ で新単語が出現。
  2. n 番目に単語 j の出現する確率 $P_{n,j}$ は、$n-1$ 番目までの単語 j の出現頻度 $s_{n-1,j}$ を用いて
  $$P_{n,j} = \frac{s_{n-1,j}^{\beta}}{\sum_k s_{n-1,k}^{\beta}} \tag{19}$$
  3. そのとき、β の値は、$n-1$ 番目に出現した単語のランク r_{n-1} と切替ランク r_c の関係から
  $$\beta = 1 + \lambda \left[\frac{r_{n-1}}{V_{n-1}} - \frac{r_c}{V_{Max}} \right] \tag{20}$$
- 新単語発生により語彙数 V_n が総語彙数 V_{Max} に達したとき試行終了

ここで、V_s に達するまでの異なる単語の出現と達した後の一定確率での新単語発生は 3.3.3 節で調べた新単語発生確率の挙動を参考に、初期単語数の数だけ新単語を発生させてから既出単語を再出現させた。式 (20) は $\frac{r_{n-1}}{V_{n-1}}$ は $n-1$ 番目に出現するランク数列 r_{n-1} をその時の総語彙数 V_{n-1} により相対化されたランクと $\frac{r_c}{V_{Max}}$ も同様に相対化された切替ランクの差を取った。$\frac{r_{n-1}}{V_{n-1}} > \frac{r_c}{V_{Max}}$ のときは一つ前の単語が上位単語であることから次に来る単語のランクはランダムに選ばれるように $\beta < 1$ となるように、$\frac{r_{n-1}}{V_{n-1}} < \frac{r_c}{V_{Max}}$ のときは一つ前の単語が下位単語であることから次に上位単語が選ばれるように $\beta > 1$ となるように定義した。設定理由については後述するが、新単語発生確率 γ=0.16、総語彙数 V_{max}=5000、初期単語数 $V_s = 30$、切替ランク $r_c = 300$、確率変動パラメータ λ=0.20 を基本パラメータとして、そのランク数列による HDM と言語テキストの HDM の比較を DC 平面上で行った。乱数による偏りをなくすために乱数の種 10 個分の平均を取りながら、新単語発生確率 γ=0.10,0.15,0.20,0.25,0.30,0.35,0.40（図 25）、初期単語数 V_s=10,30,50,100,300,900,1000,2000（図 26）、切替ランク r_c=10,30,50,100,300,900,1000,2000（図 27）、確率変動パラメータ λ=0.00,0.05,0.10,0.15,0.20,0.25,0.30（図 28）をそれぞれ変動させていったところ、どのパラメータにおいても D は大きく、C も小さいので 10 言語テキストの HDM の特徴を再現することはできなかった。

次に、先ほど使用した各パラメータを変動させていった場合のデータに対するランクサイズをプロットした。図 31、32 から、切替ランク r_c または確率変動パラメータ λ を変動させたデータのランクサイズプロットはべき乗則に乗っており、かつ、その値の違いがべき乗則の成立に与える影響を見れらない。しかし、新単語発生確率 γ または初期単語数 V_s を変動させていった場合のデータにおいてはその値が大きくなるにつれてべき乗則は大きく外れるようなプロットになっていることがわかる（図 29、30）。

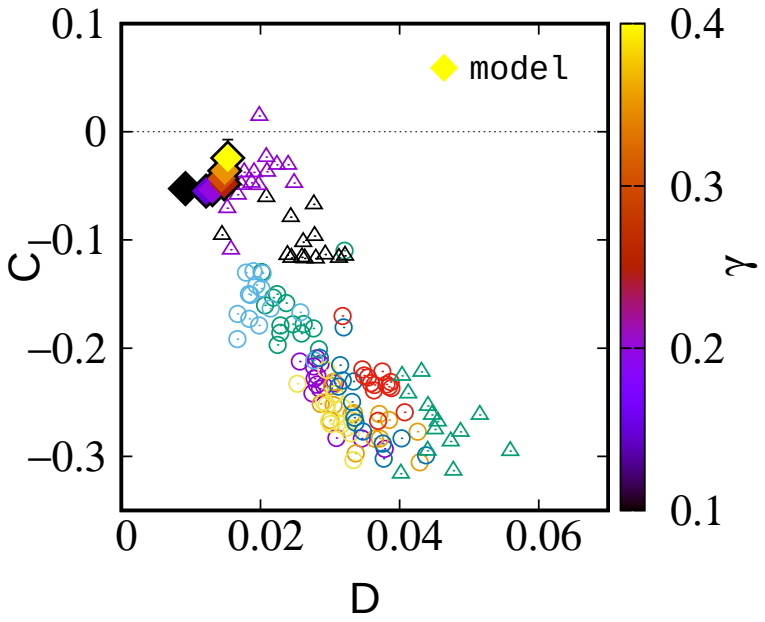

図 25: 重み線形切替 Yule モデルによる HDM への新単語発生確率 γ の依存性。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。カラーバーが新単語発生確率 γ。

4.3 重み二値切替 Yule モデル

モデルのパラメータを以下のように定義した。β の変化量は確率変動パラメータ Δ により一定とした。以下で重み二値切替 Yule モデルについてまとめた。

パラメータ	記号	説明
新単語発生確率	γ	既出でない全く新しい単語が発生する確率
総語彙数	V_{Max}	文章内に出現する単語の種類数
初期単語数	V_s	Yule 過程の開始までに既に一度ずつ発生していた単語
乱数の種	S	プログラムの乱数発生に用いた整数
切替ランク	r_c	頻出単語とそうでない単語を区別するためのランク
確率変動パラメータ	Δ	切替ランクに応じて拡張 Yule 過程の振る舞いを変更する変数

表 5: モデルのパラメータ。

- 語数が V_s になるまで、異なる単語が 1 語ずつ出現。
- 語数が V_s に達してからは拡張 Yule 過程に従って、単語が出現。
  1. 一定の確率 γ で新単語が出現。

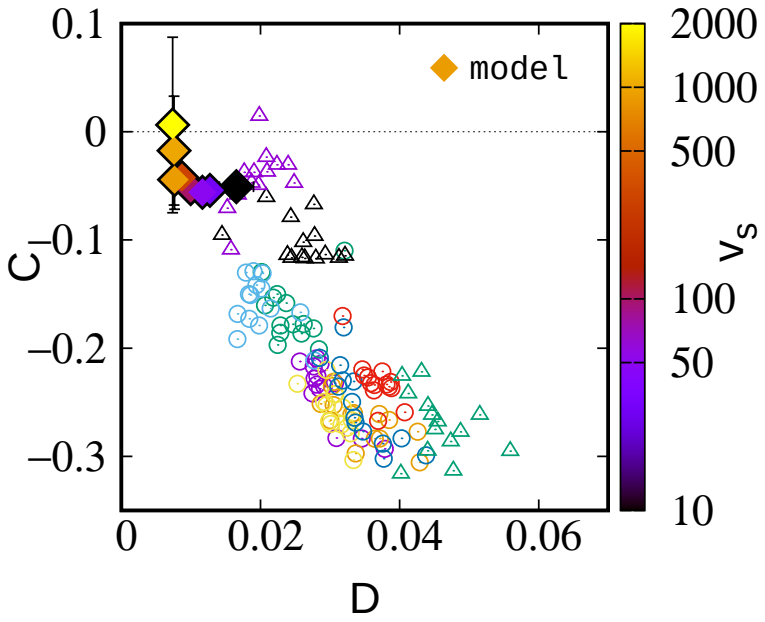

図 26: 重み線形切替 Yule モデルによる HDM への初期単語数 V_s の依存性。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。カラーバーが初期単語数 V_s。

2. n 番目に単語 j の出現する確率 $P_{n,j}$ は、$n-1$ 番目までの単語 j の出現頻度 $s_{n-1,j}$ を用いて

$$P_{n,j} = \frac{s_{n-1,j}^{\beta}}{\sum_k s_{n-1,k}^{\beta}} \tag{21}$$

3. そのとき、β の値は、$n-1$ 番目に出現した単語のランク r_{n-1} と切替ランク r_c の関係から

$$\begin{cases} \frac{r_{n-1}}{V_{n-1}} < \frac{r_c}{V_{Max}} \text{のとき} & \beta = 1+\Delta \\ \frac{r_{n-1}}{V_{n-1}} \geq \frac{r_c}{V_{Max}} \text{のとき} & \beta = 1-\Delta \end{cases}$$

- 新単語発生により語彙数 V_n が総語彙数 V_{Max} に達したとき試行終了

重み二値切替 Yule モデルにより生成されたランク数列と言語テキストの比較を DC 平面上で行った。Δ と S 以外のパラメータを固定した。新単語発生確率 γ は言語テキストの総語彙数（5000 単語）を総単語数でそれぞれ割ることで簡易的に推定（$0.1 \leq \gamma \leq 0.4$）し、その中から γ=0.16 で固定した。V_{Max} は実データとの比較を行うため 5000 で固定した。V_s は Yule 過程においてべき乗則を成立させるためには 1 でも十分であるが、実際のテキストでは一文目で同じ単語が使われることはほとんどないので 10 ～ 40 程度の新単語

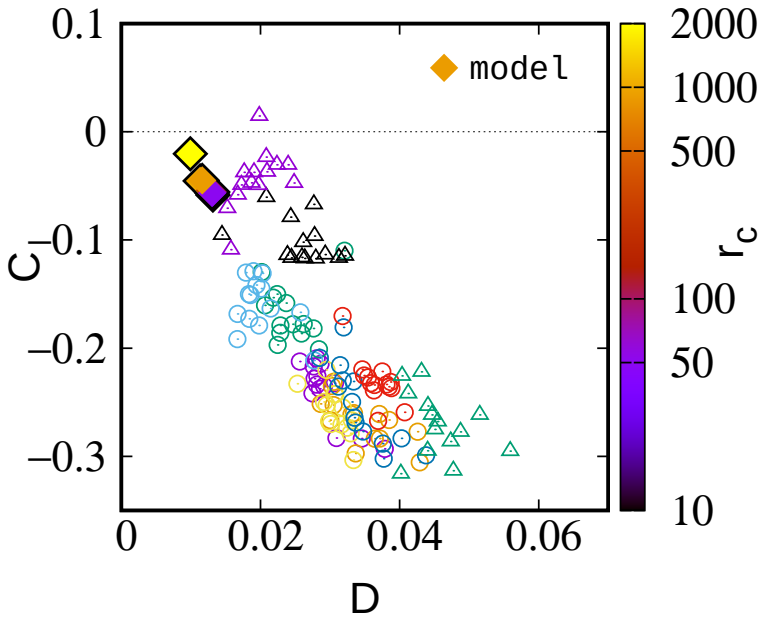

図 27: 重み線形切替 Yule モデルによる HDM への切替ランク r_c の依存性。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。カラーバーが切替ランク r_c。

が出現してから Yule 過程が開始されるとして 30 で固定した。実際に V_s が 1 と 10 以上 100 以下程度の HDM を比較したが後者の方が実データの特徴をよく再現していた。r_c はその値が大きいほど実データの HDM カラースケール $B(i,j) - \tilde{B}(i,j)$ の最大値と最小値を小さくなる傾向があり、適当な値として 300 で固定した。乱数の種 10 個分の平均をとることで乱数発生による偏りを小さくした。Δ を 0 から 0.3 まで 0.05 ずつ大きくしていったところ、言語テキストのクラスタに沿うような形で C は小さく、D は大きくなっていった（図 33）。この結果により、確率変動パラメータ Δ によるデルタモデルの言語テキスト HDM の再現性を評価できた。

他のパラメータの影響による結果を以下に示す。新単語発生確率 γ=0.10,0.15,0.20,0.25,0.30,0.35,0.40（図 25）、初期単語数 V_s=10,30,50,100,300,900,1000,2000（図 26）、切替ランク r_c=10,30,50,100,300,900,1000,2000（図 27）をそれぞれ変動させていったが重み線形切替 Yule モデルに比べて言語クラスタに近いところには位置するものの確率変動パラメータ Δ ほどの再現性がよくないことがこの結果からもよく分かる。

重み線形切替 Yule モデルと同様にそれぞれのパラメータがランクサイズプロットに与える影響を調べた。べき乗則に従うかどうかは先ほどと同様に、切替ランク r_c または確率変動パラメータ δ を変動させていった場合（図 39、40）と新単語発生確率 γ または初期単語数 V_s を変動させていった場合（図 37、38）の二つに分かれることが分かる。これまで

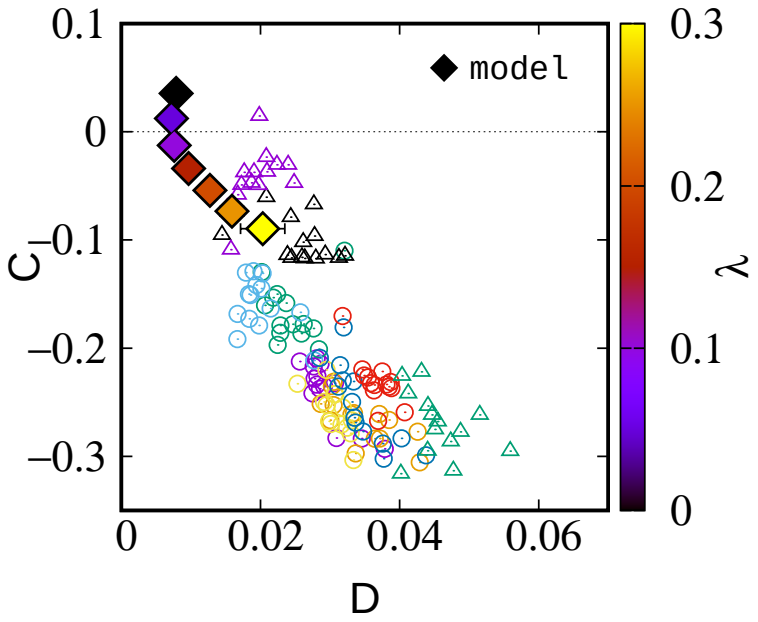

図 28: 重み線形切替 Yule モデルによる HDM への確率変動パラメータ λ の依存性。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。カラーバーが確率変動パラメータ λ。

の結果から、重み二値切替 Yule モデルにおいてべき乗則に従うことと 10 言語 HDM の特徴のどちらもランク数列に反映できる適当なパラメータは確率変動パラメータ Δ であると言える。

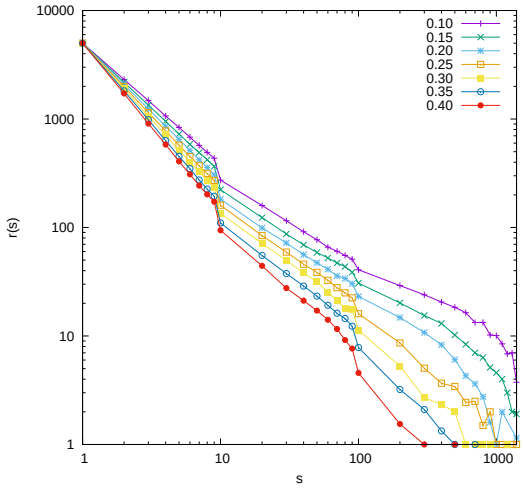

図 29: 重み線形切替 Yule モデルによるランクサイズプロットへの新単語発生確率 γ の依存性。

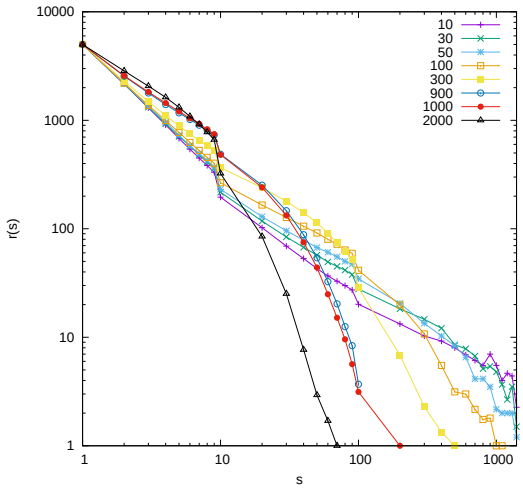

図 30: 重み線形切替 Yule モデルによるランクサイズプロットへの初期単語数 V_s の依存性。

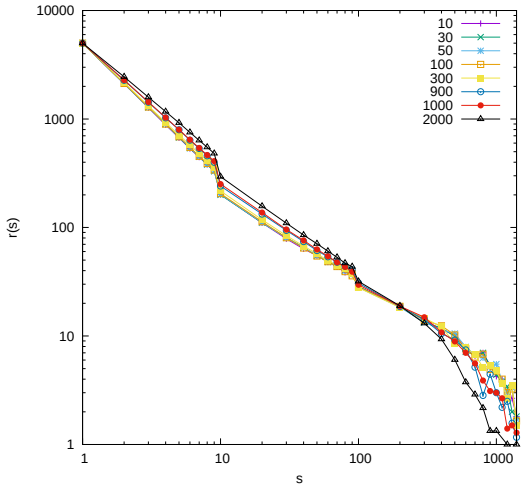

図 31: 重み線形切替 Yule モデルによるランクサイズプロットへの切替ランク r_c の依存性。

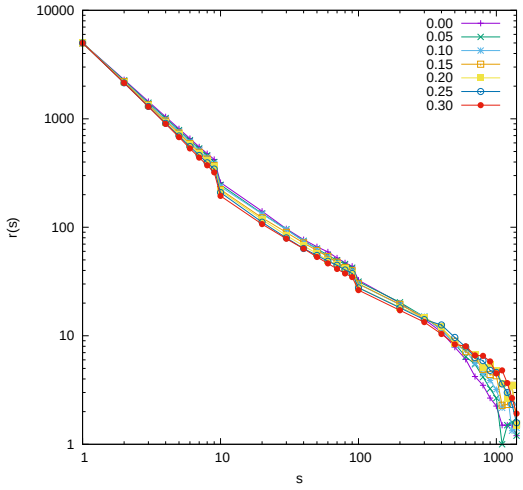

図 32: 重み線形切替 Yule モデルによるランクサイズプロットへの確率変動パラメータ λ の依存性。

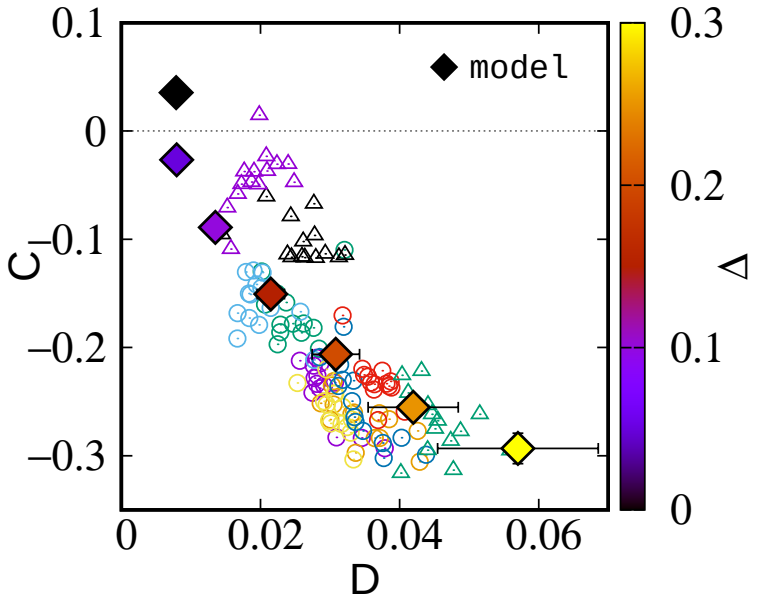

図 33: 重み二値切替 Yule モデルに対する確率変動パラメータ Δ の依存性。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。カラーバーが確率変動パラメータ Δ。

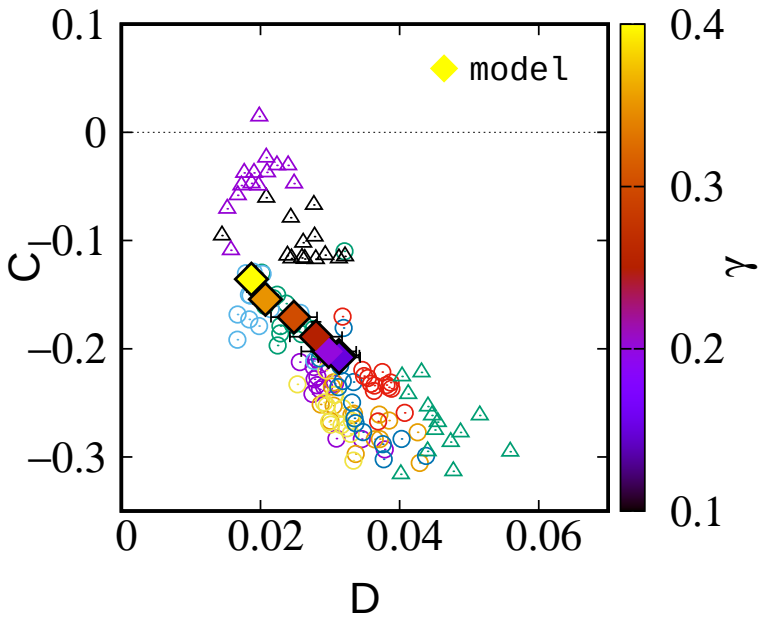

図 34: 重み二値切替 Yule モデルに対する新単語発生確率 γ の依存性。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。カラーバーが新単語発生確率 γ。

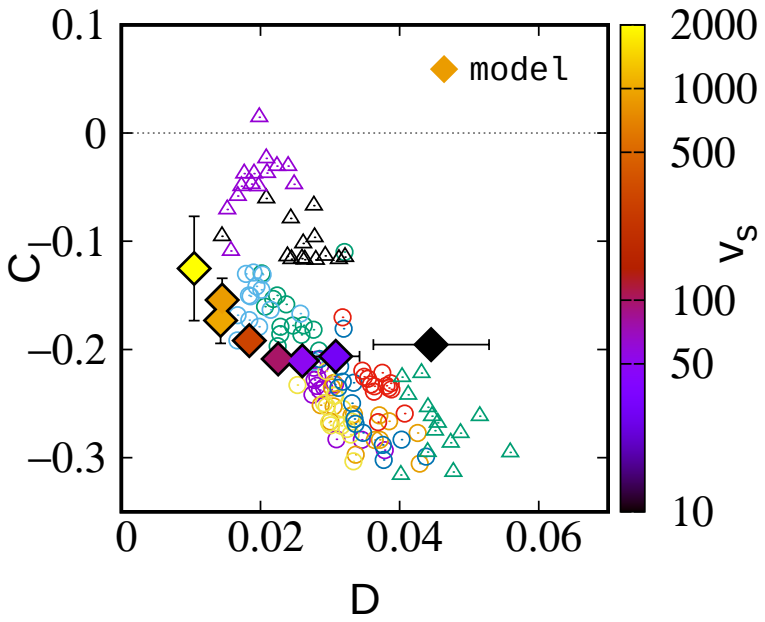

図 35: 重み二値切替 Yule モデルに対する初期単語数 V_s の依存性。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。カラーバーが初期単語数 V_s。

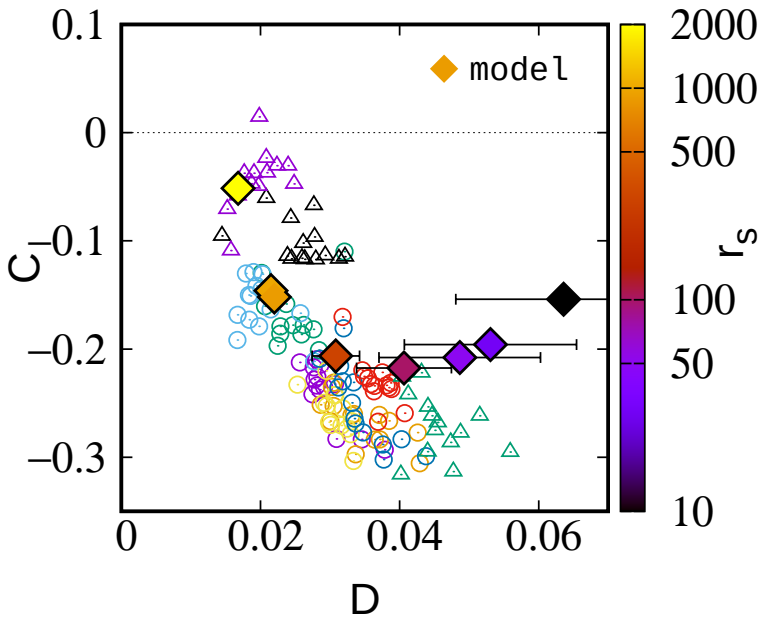

図 36: 重み二値切替 Yule モデルに対する切替ランク r_c の依存性。横軸がデータ間のユークリッド距離 D、縦軸が相関係数 C。カラーバーが切替ランク r_c。

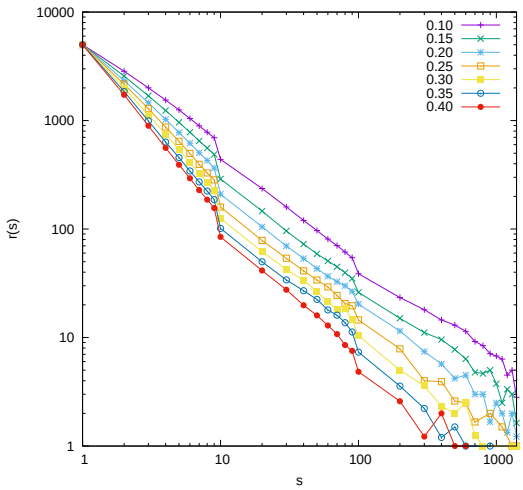

図 37: 重み二値切替 Yule モデルによるランクサイズプロットへの新単語発生確率 γ の依存性。

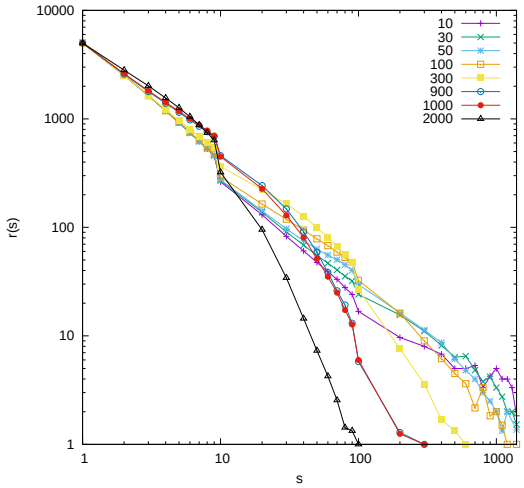

図 38: 重み二値切替 Yule モデルによるランクサイズプロットへの初期単語数 V_s の依存性。

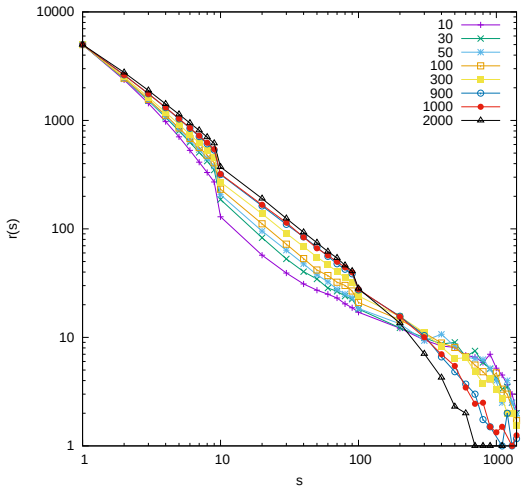

図 39: 重み二値切替 Yule モデルによるランクサイズプロットへの切替ランク r_c の依存性。

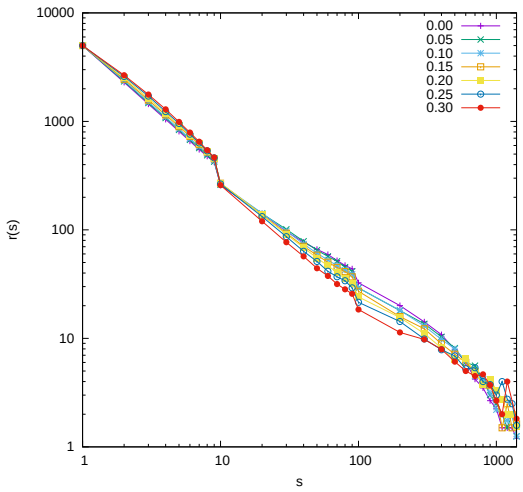

図 40: 重み二値切替 Yule モデルによるランクサイズプロットへの確率変動パラメータ Δ の依存性。

5　英語テキスト集団おける単語の非一様性と Zipf 則

75 冊の英語テキストを一つの集団とみなして単語数別の Zipf 則とそれぞれの単語の非一様性を調べた。また、その結果とこれまでの研究を参考に人間の単語の選択に関する考察を行う。（テキスト集団に対する性質としてよく知られている Heaps 則の成立については付録 A.7 に示している。）

5.1　各テキストに対する単語の非一様性

総単語数や総語彙数に関係なく英語テキストを 75 冊用意し、一つの集団とみなして解析を行った。英語テキスト集団の総単語数 $W_{all} = 10460187$、総語彙数 $V_{all} = 103549$ である。Kamada らの日本人の姓名の分布に対して都道府県別の非一様性を参考に [13]、英語テキスト集団内における各テキストへの単語の分布の非一様性を調べた。そのため、再び KL 情報量 D_i を用いて各単語 i のテキスト別 j の分布の偏りを求めた。テキスト j に出現する単語 i のサイズを $s_{i,j}$ とすると単語 i に対する KL 情報量 $D_w(i)$ は

$$D_w(i) = \sum_{j=1}^{75} \frac{s_{i,j}}{\sum_j s_{i,j}} \cdot \ln\{\frac{s_{i,j}}{\sum_j s_{i,j}} / \frac{\sum_i s_{i,j}}{W_{all}}\} \qquad (i = 1, 2 \cdots V_{all}) \tag{22}$$

となる。

初めに各単語の KL 情報量 $D_w(i)$ とそれに対する確率密度関数 $p(D_w(i))$ とテキスト集団内における単語 i の合計サイズ S_i に対する確率密度関数 $p(S_i)$ を調べた（図 41,42）。双方とも両対数プロットに対して線形になっていることからべき乗則に従っている、つまり英語テキスト集団内における単語のサイズとその非一様性に対しても Zipf 則が成立していることが分かる。これは Kamada らの日本人の姓名の分布においても同様の傾向が見られている。

次に横軸に KL 情報量 $D_w(i)$、縦軸にテキスト集団全体における単語 i のサイズ S_i をプロットし、サイズに対する非一様性を調べた（図 43）。カラースケールで KL 情報量に対する確率密度関数 $p(D_w(i))$ を表すことで同程度の KL 情報量を持つ単語の分布も示した。カラースケールにより KL 情報量 $D_w(i)$ に対するサイズ S_i も Zipf 則に従っていることが分かる。$D_w(i)$ も S_i も共に大きい所に分布する単語、つまり特定の本に集中して出現する単語には固有名詞、特に人名（その本の登場人物）が見られた。また、$D_w(i)$ が小さく S_i が非常に大きい単語、つまりすべての本に対して共通にかつ頻繁に出現する単語には the,of,can,it などの前置詞や助動詞、代名詞などが多く見られた。これはリターンマップ解析において下位単語と上位単語に分類された単語の傾向と同じである。ランク数列生成モデルにおいて上位ランクと下位ランクの単語の切替が HDM の再現に大きな役割を果たしたことも踏まえて上位単語と下位単語の間には性質的な違いがあることが示唆される。

5.2　英語テキスト集団の単語数別の Zipf 則

次に、テキスト集団全体に対するランクサイズ分布を考えた。全体（$W_{all} = 10460187$、$V_{all} = 105183$）に対するべき乗則の成立は図 44 から見て取れるが、途中で折れ曲がってい

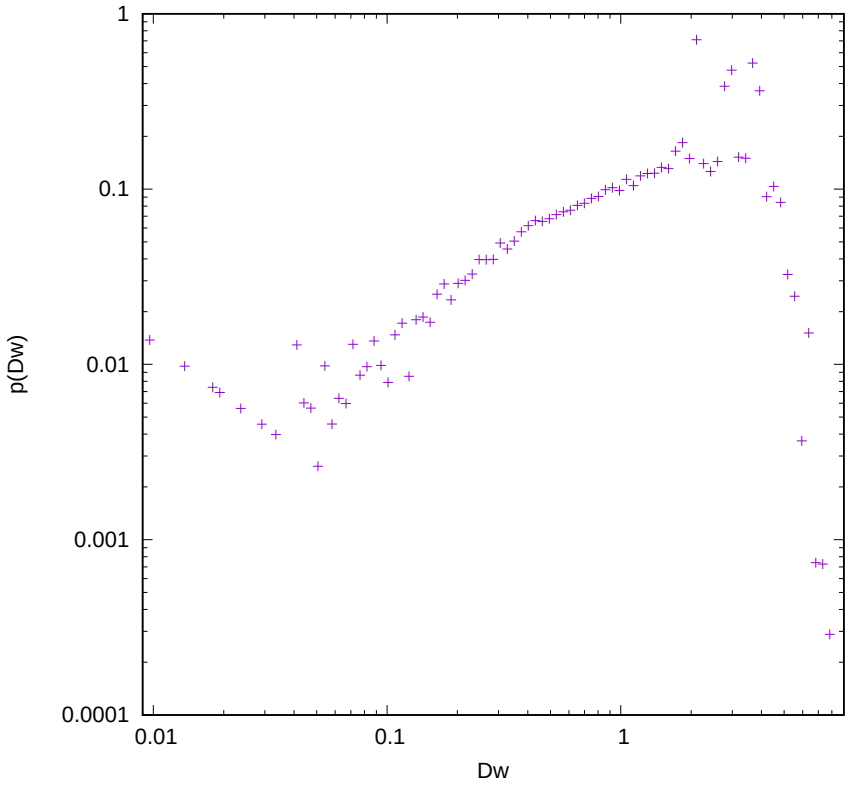

図 41: 英語テキスト集団の各単語に対する KL 情報量 $D_w(i)$ と確率密度関数 $p(D_w(i))$。

る部分があり二段階のべき乗則になっているようにも見られる。解析に使用した英語テキスト 75 冊個別のランクサイズプロット（図 45）では、折れ曲がりが確認できるテキストもそうでないテキストもあったので単語数の違いが二段階のべき乗則が生じるのではないかと考えた。そこで集団からランダムに単語を抽出し、単語数別のランクサイズをプロットした。図 46 をみると、やはり語数の多いランクサイズプロットでは折れ曲がりが確認できる。図 47 にスケーリングしたプロットを示したが総単語数が増えるにつれて折れ曲がりの度合いが大きくなることとその位置が総単語数によらないことが見て取れる。この結果から、単語数が増えるにつれて単純な Zipf 則が成立しなくなるのではないか、また二段階のべき乗則により単語はサイズに応じて二つのグループ、つまり上位単語と下位単語に分けられるのではないかと考察した。

5.3 二辞書モデルの提案による単語選択過程の解釈

リターンマップ解析やモデルによる再現の結果から単語をグループに分けるのは出現頻度（サイズ）ではないか、つまり単語は頻出単語と珍しい単語のグループに分けることができると考えた。そうすると、図 47 の折れ曲がり位置が実際の言語における切替ランクと

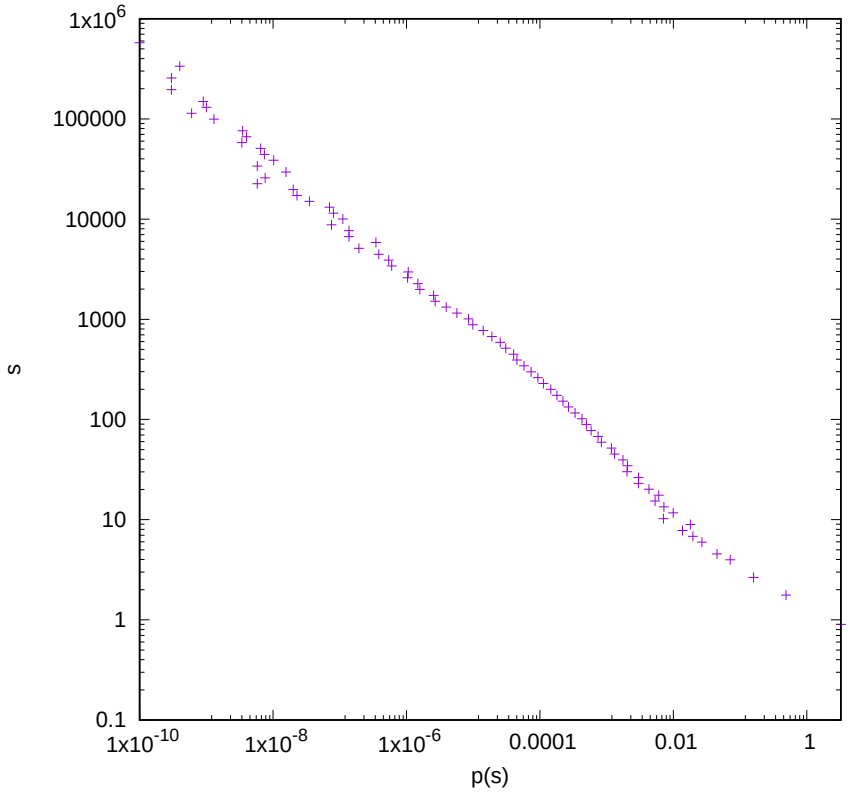

図 42: 英語テキスト集団の単語 i の合計サイズ S_i と確率密度関数 $p(S_i)$。

して考えることができる。ここで、リターンマップ解析に用いた 14 冊のデータの平均値から切替ランクを推定した。折れ曲がり位置をスケーリングサイズ $\frac{S_i}{\sqrt{W}} = 0.1$ として (11) 式に代入すると切替ランク r_c は以下のように求められる。

$$r_c = V(0.1 \cdot \sqrt{W})^{-(\alpha-1)} \tag{23}$$

(24) 式に 14 冊のデータを平均して求めた語彙数 $V = 5000$ に対する総単語数 $W = 45588.2$、2.2 節で求めた修正 Zipf 指数 $\alpha \simeq 2.00$ を用いると $r_c \simeq 234.2$ 位と求められた。重み二値切替 Yule モデルにおいて $r_c = 300$ 位により HDM を再現することができたことも踏まえると、英語テキストにおいて語彙数が $V = 5000$ の場合には切替ランクが 200 ∼ 300 位程度であると推定でき、このランクより上位のランクの単語と下位のランクの単語での振る舞いの違いがこれまでの結果から示唆された。

以上より人間が文章を書く際の単語選択に対して「二辞書モデル」を提案したい。二辞書モデルは以下のように考えられる。

- 人間は文章を書く際に仮想の 2 冊の辞書「脳内辞書」「外部辞書」から単語を選択する。

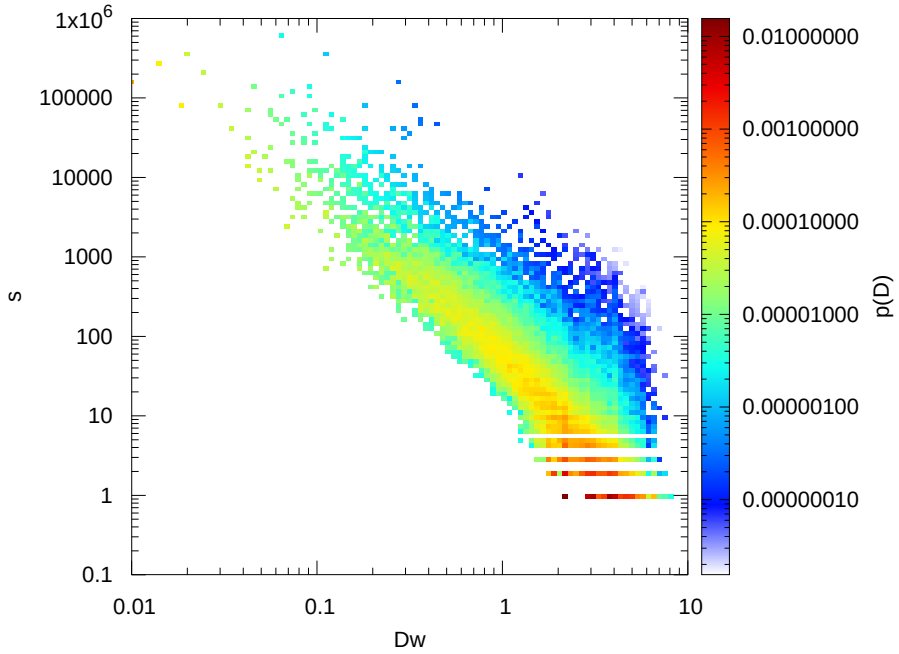

図 43: 英語テキスト集団の各単語の非一様性。横軸に KL 情報量 $D_w(i)$、縦軸にサイズ s_i、カラースケールに KL 情報量に対する確率密度関数 $p(D_w(i))$ を取っている。

- 脳内辞書には頻出単語（切替ランクよりも上位ランクの単語、主に前置詞や助動詞、代名詞など）が属しており、辞書のサイズは有限である。
- 外部辞書には珍しい単語（切替ランクよりも下位ランクの単語、脳内辞書以外の単語）が属しており、脳内辞書に比べて非常に大きいサイズを持つ。

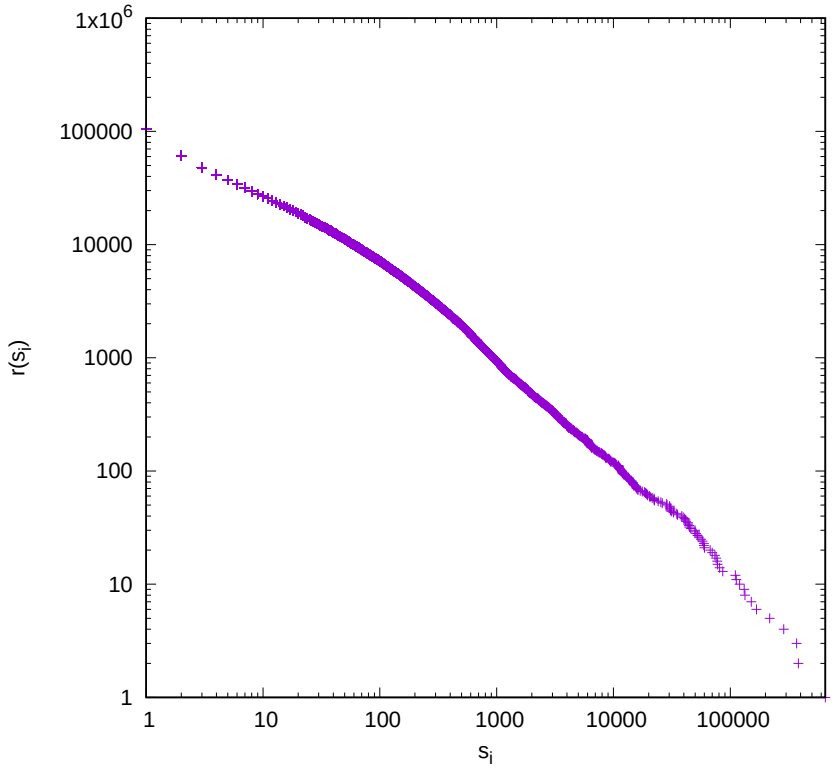

図 44: 英語テキスト集団全体のランクサイズプロット。

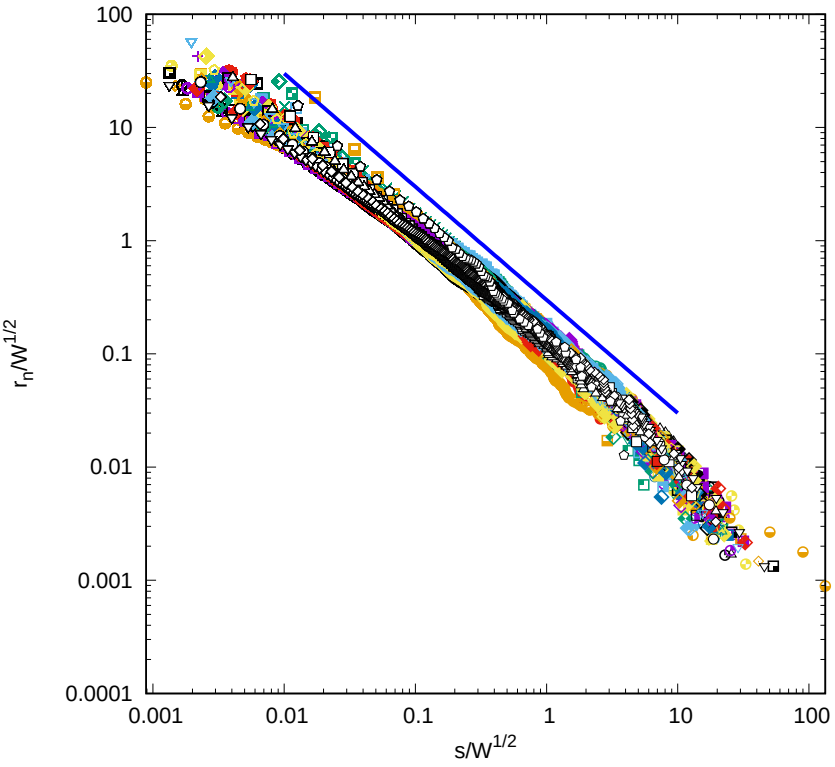

図 45: 英語テキスト集団内の個別のテキストにおけるランクサイズプロット。傾きが -1 の直線との比較。

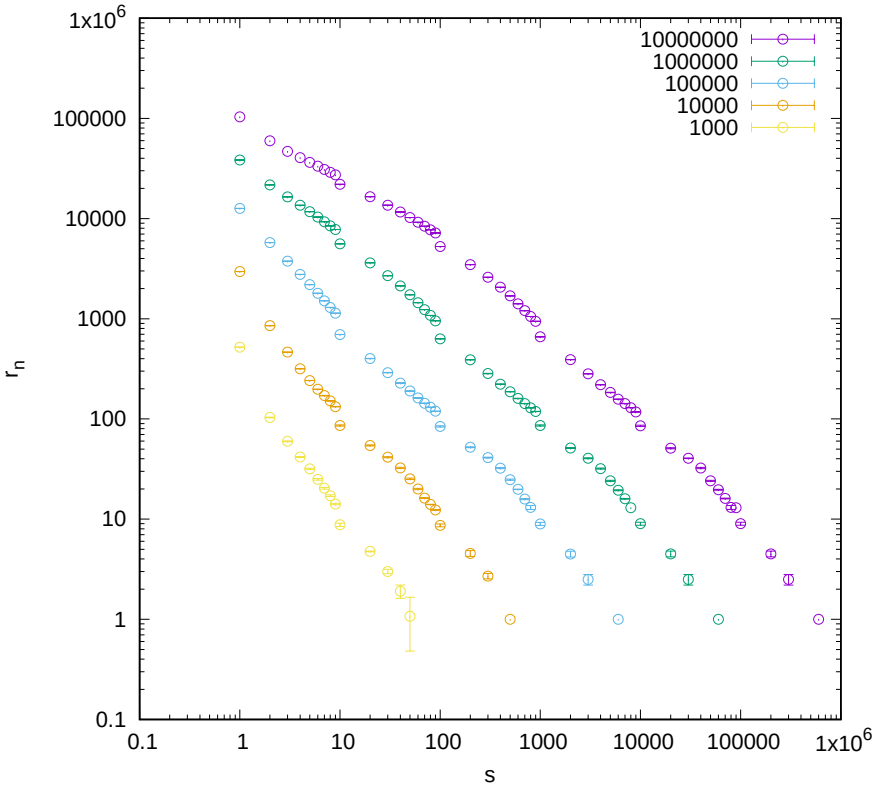

図 46: 英語テキスト集団の語数別の平均ランクサイズプロット。

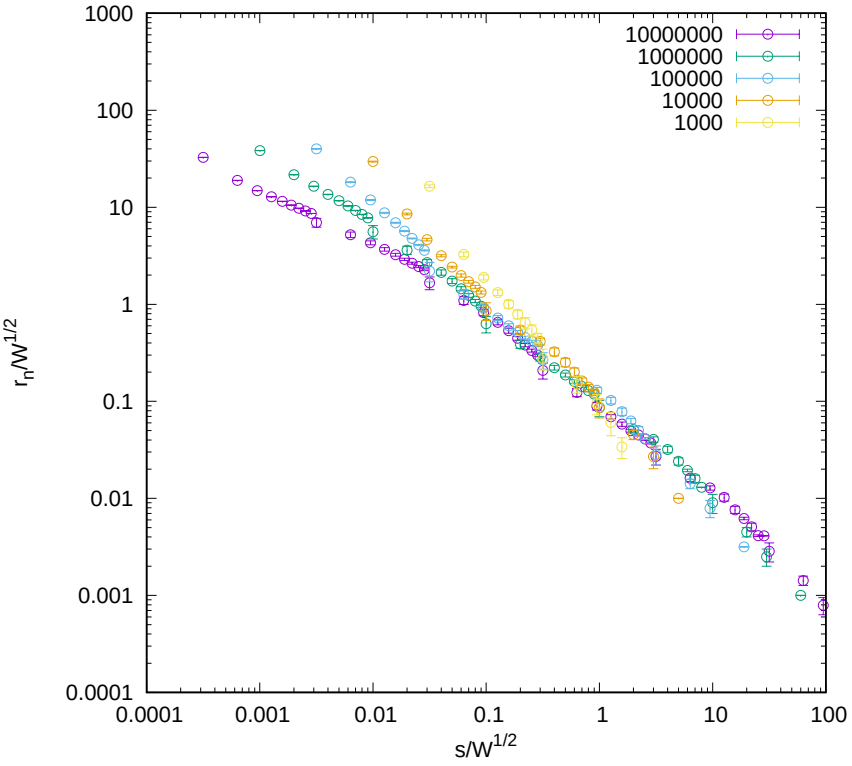

図 47: スケーリングした英語テキスト集団の語数別の平均ランクサイズプロット。

6　結論

人間の精神活動の産物である文章に対して数理的な解析を行った。先行研究の中で最も有名なZipf則は文章内の各単語の出現頻度(サイズ)とそのサイズに対して付けられたランクの間に見られるべき乗則であるが、本研究では初めにZipf則の成立をProject Gutenbergから無作為に抽出した英語テキスト14冊を使って検証した。さらに、Newmanによる先行研究ではべき指数を推定する式があったが、よいフィッティングにならなかったのでオリジナルの方法でその式を修正した。その結果、ランクサイズプロットをよくフィッティングをすることができた。

次に、分布則であるZipf則には見られない単語の並びに着目した。Zipf則で用いたランクを各単語に割り当てることによってランク数列に変換し数理的な解析を可能にした。特に隣接単語間の相関が大きいと仮定して、リターンマップを用いて隣接単語間の関係を視覚化した。Zipf則の検証に用いた英語テキスト14冊分に語彙数を5000で制限した後に、同じ単語列をランダムに再配置したランク数列（サロゲートデータ）との相対頻度の差をとり、両対数二次元ヒストグラムリターンマップ（HDM）で表したところ、よく似た特徴が見られた。これは言語全体に共通する特徴なのか、それとも言語ごとに異なる特徴を持つのかを確かめるためにオランダ語・ドイツ語・フランス語・フィンランド語テキストをProject Gutenbergから、日本語テキストを青空文庫からそれぞれ14冊ずつ抽出した。6言語84冊分のHDMを比較したところ、言語全体に共通する特徴と各言語間に見られる共通な特徴の双方が確認された。他にもHDMやランク数列に対して、階層的クラスタ解析やKullback-Leibler情報量、新単語発生確率の推定などの様々な手法で解析を行ったところHDMと同様に全体と各言語に対する共通な特徴が共に確認された。HDMの相関係数Cと元データとサロゲートデータのユークリッド距離Dという二つのオーダーパラメータを設定し、HDMの特徴を再び定量化しDC平面で比較したところ（i）ほぼ全ての文章の相関係数Cが負の値を示すことと、（ii）分布が言語ごとに偏っていることが分かった。（i）は上位ランクと下位ランクの単語が交互に並んでいるという傾向が言語によらない普遍的な性質であることを示している。また、（ii）より同一言語で書かれた文章に共通する言語の特徴が存在することも示唆される。さらに、DC平面上でクラスタ同士の分布が近い言語と言語類型論による分類に類似性が見られた。

次に、言語ランク数列の特徴はランクがサイズに対してべき乗則に従うこと(Zipf則)と、リターンマップ解析の結果から上位ランク単語と下位ランク単語の並びには相関があることであると仮定した。この二つの特徴を再現する数理モデルを以下のように構築した。まず、べき乗則を再現するためにYule過程すなわち一定の割合で新単語が出現し、既出単語は自身の出現回数に比例した割合で再出現するというモデルをベースとした。我々はさらに上位ランクと下位ランクを分ける切替ランクを定義し、既出単語の再出現確率を隣接単語のランクに応じて重み付けを行った。重み付けの度合いも隣接ランクに応じて線形に変動するモデル（重み線形切替Yuleモデル）と度合いが一定であるモデル（重み二値切替Yuleモデル）の二つを提案した。モデルによって生成されたランク数列に対するリターンマップ解析を行い、そのHDMをDC平面上で10言語のHDMと比較した結果、両者でべき乗則の成立は確認できたが、HDMの特徴は前者のモデルでは上手く再現することはできなかった。これに対して、後者のモデルにおいて重み付けのパラメータを変化させることで言語ごとのHDMの特徴をある程度再現することができた。これより、切替ランクが

文章を書く際に大きな役割を果たすことが示唆された。また、ランクサイズプロットを各パラメータごとに比較したが、これに対しても重み付けのパラメータはべき乗則の成立という条件を満たしていた。

これまではそれぞれの文章に対して個々の解析を行ってきたが、リターンマップ解析などから言語間に共通する特徴が示唆されたため、英語テキスト 75 冊を一つの集団とみなして解析を行った。一つ目に、再び KL 情報量を用いて単語の各テキストへの分布に対する非一様性を調べた。その結果、Zipf 則で上位単語に分類された前置詞や助動詞、代名詞などが全てのテキストに対して共通かつ頻繁に出現することと下位単語に分類された単語は局在し、主に人名などの固有名詞が固有の本に集中して出現することなどが分かった。これからも上位単語と下位単語間における性質の違いが示唆された。加えて、英語テキスト集団全体に対する Zipf 則の成立を調べたところ、折れ曲がりのある二段階のべき乗則が確認された。個別の本に対するプロットでは折れ曲がりが確認できるテキストもそうでないテキストもあったため、総単語数の増加が原因ではないかと考えた。これを確かめるために集団からランダムに単語を抽出した単語数別のランクサイズをプロットした。スケーリングにより語数別のプロットを比較したところ、語数が増えるにつれて折れ曲がりの度合いが大きくなることと折れ曲がりの位置の相対サイズが一様であることが確認できた。これから、語数が増えるにつれて単純な Zipf 則が成立しなくなることと、二段階の Zipf 則より上位単語と下位単語にはやはり性質的な違いがあることが示唆された。さらに、相対サイズと英語の実データから切替ランクを推定したところ、重み二値切替 Yule モデルで良い精度を与えたパラメータとしての切替ランクと近い値が得られた。これにより、実際の言語においても切替ランクの存在が推定された。

これまでの解析結果から、人間は文章を書く際に上位単語が属する脳内辞書と下位単語が属する外部辞書を使い分けを行なっているのではないかとという二辞書モデルを提案したい。

今後の予定としては、DC 平面上で解析する言語を増やし言語類型論との類似性を定量的に捉えること、HDM に対するさらに精度の良い解析方法の提案、再現がさらに高い新しいランク数列生成モデルの提案などがある。加えて、二辞書モデルの検証として本研究で解析した英語以外の 9 言語のテキスト集団に対する KL 情報量を用いた単語の非一様性の解析、総単語数別のランクサイズプロットの比較そして切替ランクの推定にも取り組みたい。ここで、他言語の切替ランクについては 3.3.2 節で推定した HDM の分布の違いによる切替ランクは言語によって異なる値を示したが、4.3 節の重み二値切替 Yule モデルで切替ランクを変更させていった場合に DC 平面上で大きな挙動の変化がなかったことから他言語における切替ランクも同一の相対サイズで推定できるのではないかと考えている。加えて、本研究で解析対象に用いたテキストは主に小説テキストが中心であったために作者の言語能力やテキストのジャンルによる相関も結果に影響した可能性があるので、学術論文やエッセイ、歴史書など様々なジャンル、また作者の年齢や時代背景の異なるテキストに対する解析も行なっていきたい。また、本研究において使用した言語は言語類型論において膠着語と屈折語に分類される言語のみであり、孤立語（中国語やタイ語など）は解析対象に含まれていなかったので今後は孤立語の解析対象に含め言語類型論と DC 平面上でのクラスタの分布の類似性を調べたい。また、同一言語テキスト集団における Zipf 則の折れ曲がりの度合いの変化についてもさらに総単語数を増やしていき、折れ曲がりの度合

いと総単語数の関係も定量的に捉えていきたい。将来的には、この研究を発展させていくことで隣接単語間に見られる出現頻度に依存した単語の相関や、単語の非一様性に対する解析を利用して、ヴォイニッチ手稿や古代言語などの未解読のテキストの解読などにこの研究を発展させていきたい。

謝辞

形態的言語類型論に関して助言をいただきました青山学院大学理工学部の鈴木岳人先生に厚く御礼申し上げます。また、本研究を遂行するにあたり、熱心なご指導と適切な助言を頂きました指導研究員の水口毅先生に深く感謝致します。そして、経済面や生活面を支えて下さった家族に深い敬意と感謝を示し、本論文の結びとさせていただきます。

参考文献

[1] "Modeling Statistical Properties of Written text", M. Angeles Serrano, Alessandro Flammini, Filipo Menczer, *PLoS ONE* **4**(2009), e5372.

[2] "Power laws, Pareto distributions and Zipf's law", M. E. J Newman, *Contemporary Physics* **46**(2005), pp.323-351.

[3] "Size Frequency Distribution of Japanese Given Names", Ryo Hayakawa, Yuta Fukuoka, Tsuyoshi Mizuguchi, *Journal of the Physical Society of Japan* **81**(2012) 094001

[4] "Using two-stage conditional word frequency models to model word burstiness and motivating TF-IDF", Peter Sunehag, *Journal of Machine Learning Reserach* (January 2007)

[5] "Modeling Word Burstiness Using the Dirichlet Distribution", Rasmus E. Madsen, David Kauchak, Charles Elkan, *ICML '05: Proceedings of the 22nd international conference on Machine learning* (August 2005) pp.545-552.

[6] "Long-range correlations in nucleotide sequnces", C. -K. Peng et al., *NATURE* **356**(1992), pp.168-170.

[7] "LONG RANGE CORRELATION IN HUMAN WRITINGS", A. Schenkel, J. Zhang and Y. -C. Zhang, *Fractals* **1**(1993), pp.47-57.

[8] "Complexity and human writing", Peter Kokol, Vili Podgorelec, *Complexity International* **7**(2000), pp.1-6.

[9] "Entropy and Long-Range Correlations in Literary English.", W. Ebeling, T. Poschel, *EUROPHYSICS LETTERS* **26**(1994), pp.241-246.

[10] "Hierarchinal structures induce long-range dynamical correlations in writeen texts", E. ALverez-Lacalle et al., *PNAS* **103**(2006), pp.7956-7961.

[11] "On the origin of long-range correlations in texts", Eduardo G. Altmann, Giampaolo, Mirko Degli Esposti, *PNAS* **109**(2012), pp.11582-11587.

[12] "Universal Entropy of Word Ordering Across Linguistic Families", Marcelo A. Montemmurro, Damian H. Zannette, *PLoS ONE* **6**(2011), e19875.

[13] "Heterogeneity of Japanese Names", Ryohei Kamada, Tsuyoshi Mizuguchi, *The Physical Society of Japan* **89**(2020), 074802.

[14] Project Gutenberg, `https://www.gutenberg.org`.

[15] Aozora Bunko, `https://www.aozora.gr.jp`.

A　付録

A.1　解析の開始位置を変えた場合の相関

本研究では使用するテキストを全て冒頭から解析した結果を用いたが冒頭以外から解析を開始した場合についてオーダーパラメータ D と C を用いて比較した。英語版 MobyDick（総単語数 207007）を単語数ごとに四等分、つまり単語数が 50000 語ずつで解析位置をずらしていき、それぞれ（$1/4$ = 通常の解析, $2/4$ = 単語数 50001 語目から, $3/4$ = 単語数 100001 語目から, $4/4$ = 単語数 150001 語目から）について総語彙数が 5000 になるまでの HDM を DC 平面に追加した（図 A.1）。その結果、DC 平面上で大きな変化は見られなかった。

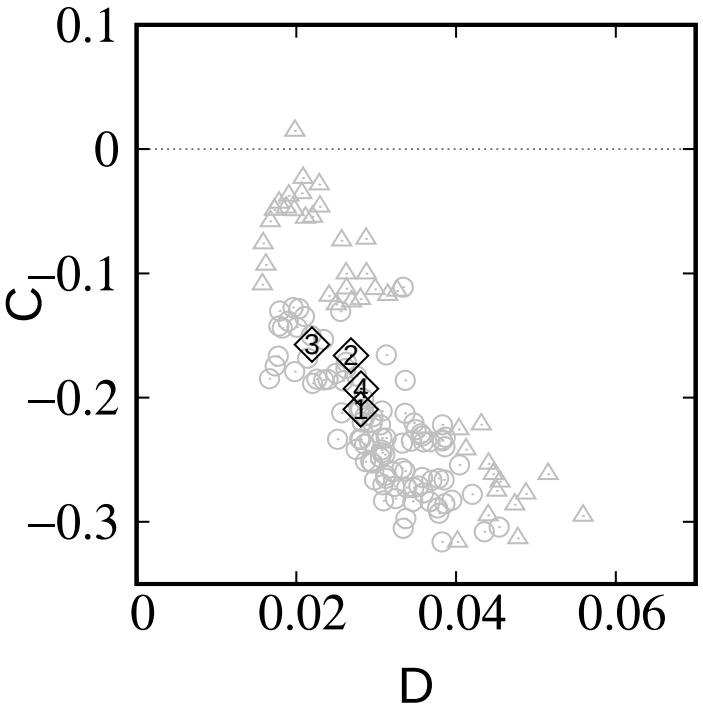

図 A.1: 解析開始位置を変えた場合の英語版 MobyDick の DC 平面。

次に、冒頭以外からテキストを使用した場合の Zipf 則の成立も調べた。英語版 MobyDick をそれぞれ $1/4, 2/4/3/4$ に分割したテキストを調べた結果として、個々の単語のランクの入れ替わりは見られたものの文章は先頭から使用せずとも成立することが分かった（図 A.2 参照）。このとき、英語版 MobyDick の文章全体 (207007 語) に対するランクサイズを測定した時の上位 10 位は the,of,and,a,to,in,that,his,it,i である。しかし、本解析では複数の文章に対してリターンマップ解析を行う際に条件を統一するために冒頭から総語彙

数 5000 語（総単語数 23983 語）までに限定した場合（1/4 テキスト）の上位 10 単語は the,and,a,of,to,in,i,his,that,he である。加えて、2/4 テキストの場合には the,of,and,to,in,a,that,his,it,is、3/4 テキストの場合には the,of,and,to,a,in,that,it,is,his、4/4 テキストの場合には the,and,of,in,a,to,it,that,his,i となっている。

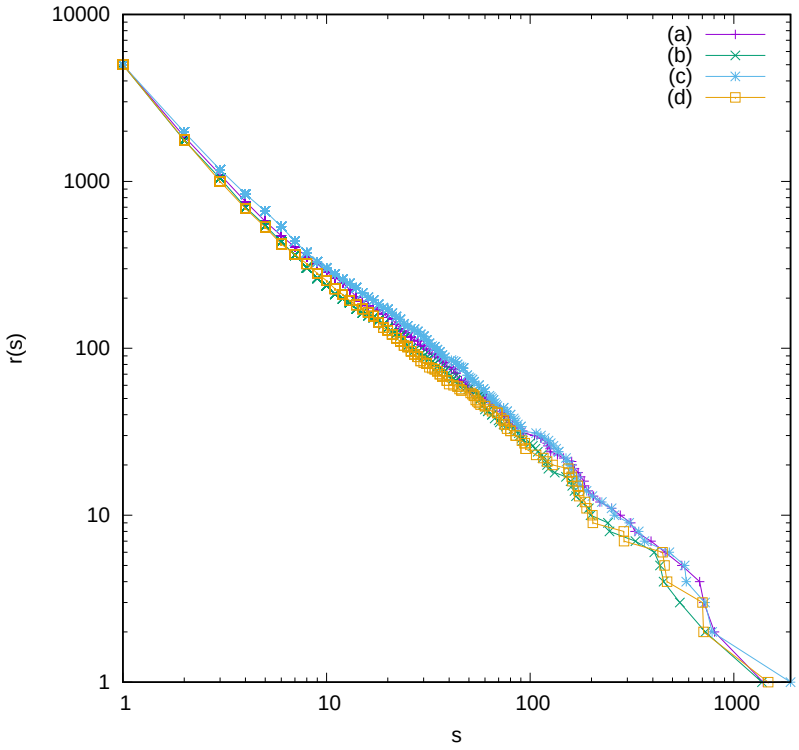

図 A.2: 解析開始位置を変えた場合の英語版 MobyDick のランクサイズプロット。(a)1/4 テキスト、(b)2/4 テキスト、(c)3/4 テキスト、(d)4/4 テキスト。

A.2 相関の単語間距離依存性

相関が最も大きいとして本研究では隣接単語（単語間距離 $k = 1$）間の関係に着目したが、単語間距離 $+k(k = 1 \sim 10)$ と徐々に大きくしていった場合の相関を調べた。図 A.3 に英語版 MobyDick の HDM のオーダパラメータ D と C の k との関係を示したところ、予想通り k が大きくなるにつれて D は小さくなり、C は大きくなっている。これはリターンマップ解析で示唆された 10 言語 HDM の特徴が小さくなっていることを示している。

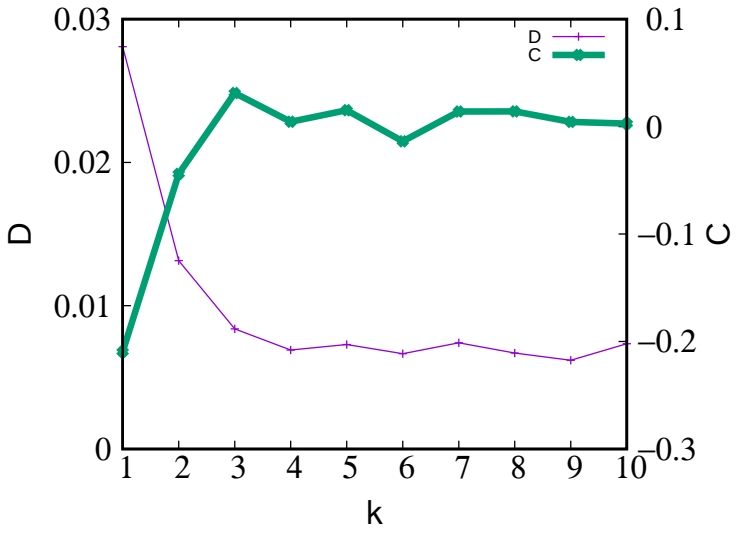

図 A.3: 英語版 MobyDick の単語間距離 k とオーダーパラメータ D と C の関係。紫線:D と k の関係。緑線:C と k の関係。

A.3　10 言語の階層的クラスタ解析

6 言語の階層的クラスタ解析の結果は図 16 で示したが後に追加した 4 言語を合わせた 10 言語の結果を図 A.4 に示した。6 言語と同様に 10 言語でも言語ごとのクラスタリングはある程度なされていた。

A.4　KL 情報量による追加 4 言語の切替ランクの推定

3.3.2 節と同様に追加 4 言語に対しても HDM の列ごとの KL 情報量を図 A.5 に示した。ポルトガル語・イタリア語・スペイン語では切替ランク r_c の推定は 30 ∼ 70 位前後であると推定できるが、ハンガリー語ではフィンランド語と同様に傾向が見えづらかった。この手法では、やはり切替ランクを決めることは難しい。

A.5　追加 4 言語の新単語発生確率

3.3.3 節と同様に追加 4 言語に対しても HDM の列ごとの KL 情報量を図 A.6 に示した。6 言語の解析から得られた新単語発生確率 $P_{nw}(n)$ は一度減少しはじめると線形に減少するという傾向は追加 4 言語にも見られることからこの性質は言語全体に共通する性質なのではないかとは考察できる。

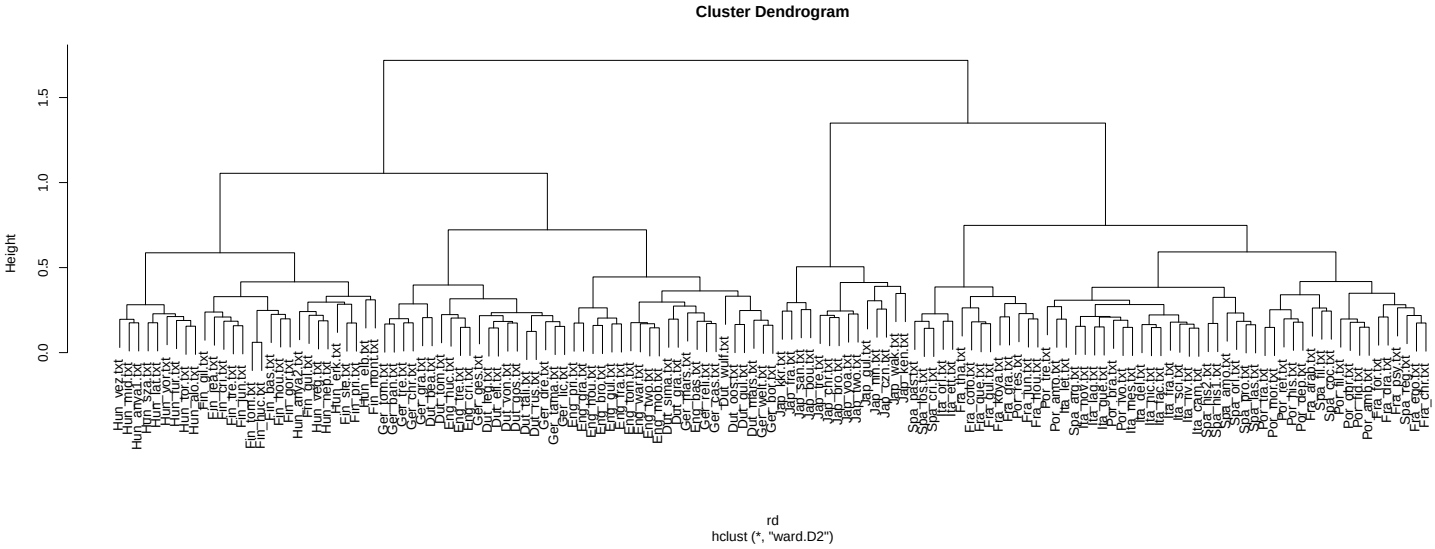

図 A.4: 距離にユークリッド距離を、距離測定法に Ward 法を用いてクラスタリングした 10 言語テキストのデンドログラム。縦軸がクラスタ間距離。

A.6　10 言語テキストのべき指数

リターンマップ解析に使用した 10 言語（英語・オランダ語・ドイツ語・フランス語・フィンランド語・ハンガリー語・イタリア語・ポルトガル語・スペイン語・日本語）テキスト 14 冊ずつ計 140 冊分の前処理だけを施したテキスト（パターン A）と、リターンマップ解析のために前処理に加えて冒頭から総語彙数が 5000 になるまでに制限したテキスト（パターン B）のべき指数を言語別の表にまとめた。表には、式 (5) から求めたべき指数（タイプ X）、2.2 節で修正したべき指数（タイプ Y）、また比較のためグラフ描画ソフト gnuplot を用いて非線型最小二乗法から求めたべき指数（タイプ Z）の 3 タイプをテキスト 2 パターン別の計 6 種類を記載している。例えば、前処理のみのテキスト（パターン A）の修正べき指数（タイプ Y）は α_{AY} のようにべき指数 α の種類は添字で区別している。

パターン A、B の両者の全てのデータにおいてタイプ X よりもタイプ Y の方がタイプ Z に近い値を取っている（図 A.7）。タイプ Z は線形なプロットに対して信頼性の高いフィッティング方法であることから、式 (5) を修正したべき指数（タイプ Y）はランクサイズプロットにおいて非常に有効な手法であると考察できる。

A.7　英語テキスト集団の Heaps 則

初めに Heaps 則はテキスト内に出現する単語の総語彙数 V と総単語数 W がべき乗則に従うという経験則である。5 章で使用した総単語数、総語彙数の異なる英語テキスト 75 冊に対して、これらを両対数プロットしてみたところ線形な形をしており Heaps 則へ従うことが確認された（図 A.8）。

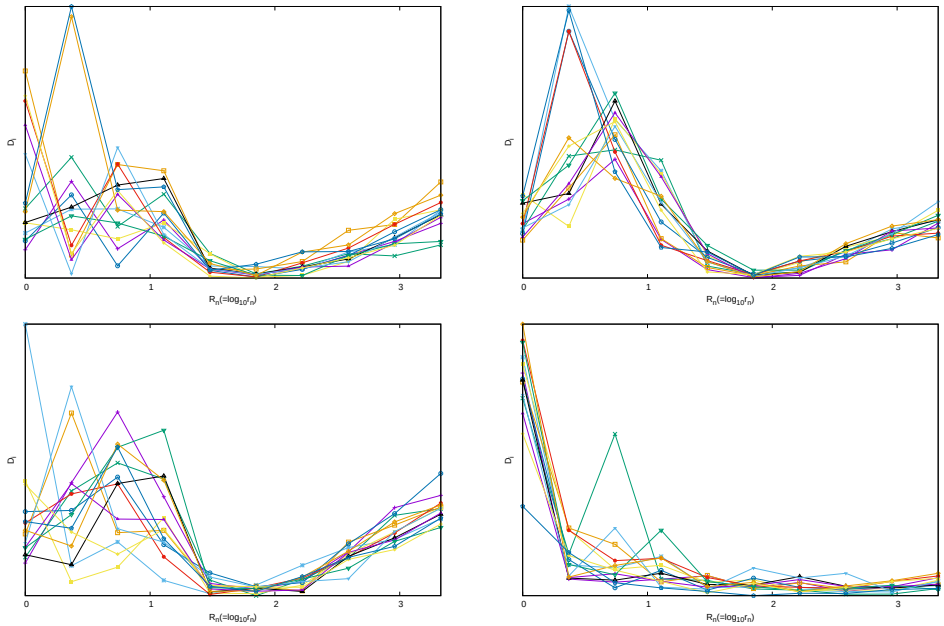

図 A.5: 追加 4 言語の KL 情報量。上列左からポルトガル語、イタリア語。下列左からスペイン語、ハンガリー語。横軸が対数ランク列 R_n、縦軸が HDM の縦列ごとの KL 情報量 D_i。

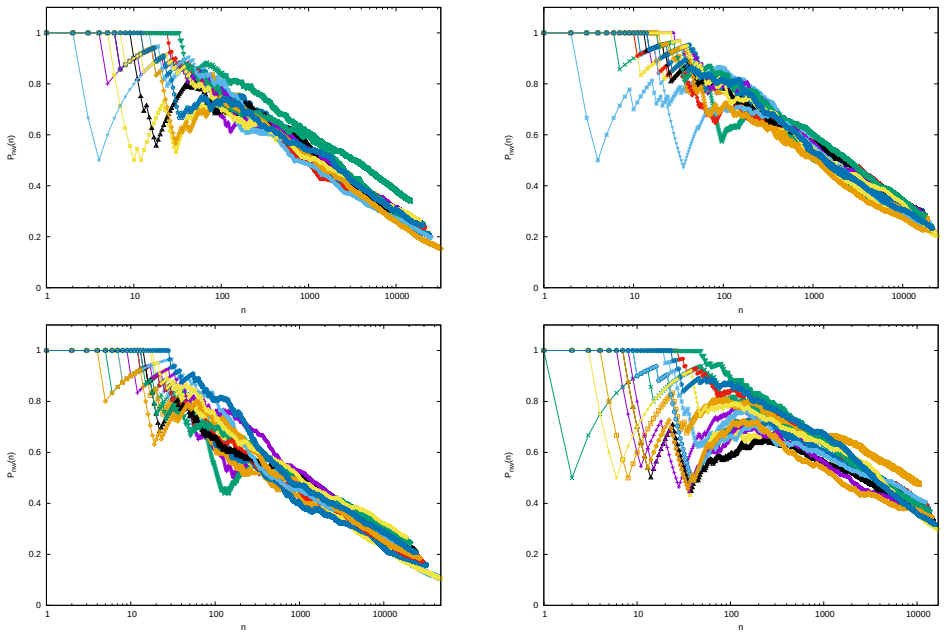

図 A.6: 追加 4 言語の新単語発生確率 $P_{nw}(n)$。上列左からポルトガル語、イタリア語。下列左からスペイン語、ハンガリー語。横軸が項番号 n、縦軸が新単語発生確率 $P_{nw}(n)$。

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	1.97	1.95	1.86	2.18	2.02	1.96
(b)	1.76	1.80	1.73	2.20	2.00	1.96
(c)	1.92	1.85	1.79	2.17	2.00	1.94
(d)	2.05	2.06	1.92	2.33	2.17	2.05
(e)	2.12	1.93	1.90	2.19	1.95	1.93
(f)	2.02	2.00	1.89	2.21	2.07	1.98
(g)	1.97	1.94	1.86	2.18	2.02	1.96
(h)	1.92	1.84	1.79	1.99	1.88	1.84
(i)	1.74	1.79	1.73	2.19	2.00	1.95
(j)	2.26	2.03	1.98	2.90	2.03	2.20
(k)	1.83	1.84	1.77	1.94	1.92	1.84
(l)	2.19	2.01	1.95	2.42	2.05	2.05
(m)	2.09	1.92	1.89	2.14	1.95	1.92
(n)	2.02	1.78	1.73	2.19	2.01	1.95

表 A.1: 英語テキスト 14 冊分のべき指数 α。(a)The hound of the Baskervilles, (b)Brothers of Kramazov, (C)Crime and Punishment, (d)Frankenstein, (e)The picture of Dorian Gray, (f)Gulliver's travels, (g)The house of the dead, (h)Adventures of Huckleberry, (i)War and Peace, (j)MobyDick, (k)Prime and prejudice, (l)The adventures of Tomsawyears, (m)Treasure Island, (n)The tale of two cities.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	2.40	1.96	1.99	2.44	1.97	2.00
(b)	2.41	2.06	2.04	2.50	2.11	2.08
(c)	2.34	2.00	2.00	2.66	2.04	2.11
(d)	2.17	1.93	1.92	2.37	1.96	1.99
(e)	2.16	1.93	1.91	2.37	1.96	1.99
(f)	2.15	1.93	1.91	2.42	2.04	2.03
(g)	2.15	1.97	1.93	2.81	2.14	2.19
(h)	2.44	2.10	2.06	2.51	2.13	2.09
(i)	2.20	2.02	1.96	2.47	2.10	2.08
(j)	2.79	2.12	2.19	2.96	2.13	2.23
(k)	2.30	1.98	1.98	2.55	2.03	2.07
(l)	2.15	2.01	1.93	2.65	2.04	2.11
(m)	2.50	1.98	2.05	2.52	1.98	2.05
(n)	2.71	2.15	2.18	3.09	2.20	2.29

表 A.2: オランダ語テキスト 14 冊分のべき指数 α。(a)Het portret van Dorian Gray, (b)De verrezen Gulliver, (c)De Lotgevallen van Tom Sawyer, (d)Beatrice, (e)Eline Vere Een Haagsche roman, (f)Gosta Berling, (g)De legende en de heldhaftige, vroolijke en roemrijke daden van Uilenspiegel en Lamme Goedzak in Vlaanderenland en elders, (h)Per luchtschip de Argonaut naar Mars, (i)De wijzen van het Oosten Brahmanisme, Boeddhisme, Chineesche philosophie, Mazdeisme, (j)Martelaren van Rusland, (k)Het boek van Siman den Javaan Een roman van rijst, dividend en menschelijkheid, (l)De Talisman of Richard Leeuwenhard in Palestina, (m)Vonken, (n)Beowulf Angelsaksisch volksepos vertaald in stafrijm en met inleiding en aanteekeningen voorzien.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	2.71	2.06	2.12	2.72	2.06	2.13
(b)	2.66	1.99	2.07	2.83	2.02	2.15
(c)	2.52	2.01	2.05	2.79	2.09	2.16
(d)	2.48	1.99	2.03	2.71	2.06	2.12
(e)	2.57	1.99	2.07	2.83	2.04	2.17
(f)	2.20	2.01	1.96	2.91	2.12	2.24
(g)	2.36	2.02	2.01	2.76	2.09	2.17
(h)	2.68	2.05	2.12	3.05	2.12	2.26
(i)	2.75	2.21	2.20	3.23	2.12	2.35
(j)	2.30	1.99	1.98	2.72	2.06	2.14
(k)	2.74	2.12	2.18	3.01	2.15	2.26
(l)	2.39	1.96	2.00	2.47	2.01	2.04
(m)	2.52	2.01	2.06	3.01	2.08	2.25
(n)	2.46	2.13	2.08	2.97	2.24	2.28

表 A.3: ドイツ語テキスト 14 冊分のべき指数 α。(a)Der Weihnachtsabend Eine Geistergeschichte, (b)Das Bildnis des Dorian Gray, (c)Die Abenteuer Tom Sawyers, (d)Die Schatzinsel: Roman, (e)Unsichtbare Bande Erzahlungen, (f)Quer Durch Borneo Ergebnisse seiner Reisen in den Jahren 1894, 1896-97 und 1898-1900; Erster Teil, (g)Die Abtissin von Castro, (h)Drei Monate Fabrikarbeiter und Handwerksbursche Eine praktische Studie, (i)Geschichte von England seit der Thronbesteigung Jakob's des Zweiten Zweiter Band, (j)Lichtenstein, (k)Die Weltensegler. Drei Jahre auf dem Mars. 1910(l)Die Religion innerhalb der Grenzen der blo β en Vernunft Text der Ausgabe 1793, mit Beifugung der Abweichungen der Ausgabe 1794, (m)Die Klerisei, (n)Der Weltkrieg, II. Band Vom Kriegsausbruch bis zum uneingeschrankten U-Bootkrieg.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	2.65	2.03	2.11	2.78	2.06	2.16
(b)	2.47	2.03	2.05	2.76	2.04	2.14
(c)	2.36	2.09	2.04	2.91	2.11	2.23
(d)	2.30	2.03	2.00	2.95	2.08	2.23
(e)	2.37	2.01	2.01	2.49	2.05	2.06
(f)	2.19	2.08	1.98	2.75	2.15	2.19
(g)	2.24	2.04	1.98	2.62	2.21	2.16
(h)	2.31	2.12	2.12	2.86	2.15	2.22
(i)	2.32	2.10	2.04	3.53	2.24	2.42
(j)	2.34	2.09	2.04	2.90	2.19	2.25
(k)	2.02	1.92	1.86	2.29	2.02	1.99
(l)	2.09	1.98	1.91	2.91	2.07	2.23
(m)	2.44	2.06	2.05	2.62	2.10	2.13
(n)	2.02	1.96	1.93	2.74	2.05	2.14

表 A.4: フランス語テキスト 14 冊分のべき指数 α。(a)Cantique de Noel, (b)Le portrait de Dorian Gray, (c)Les voyages de Gulliver, (d)Souvenirs de la maison des morts, (e)Le livre de la Jungle, (f)Vie de Christophe Colomb, (g)Psychologie de l'education, (h)Thais, (i)La Force Le Temps et la Vie, (j)Le Dragon Imperial, (k)Les mille et une nuits Tome premier, (l)Quentin Durward, (m)Souvenirs d'egotisme autobiographie et lettres inedites publiees par Casimir Stryienski, (n)La case de l'oncle Tom ou vie des negres en Amerique.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	3.06	2.18	2.28	3.64	2.30	2.48
(b)	2.69	2.09	2.13	3.71	2.35	2.50
(c)	3.13	2.17	2.29	3.23	2.23	2.33
(d)	3.54	2.34	2.46	4.20	2.41	2.65
(e)	2.92	2.21	2.26	4.07	2.39	2.61
(f)	2.85	2.11	2.20	3.51	2.31	2.44
(g)	3.60	2.26	2.45	4.24	2.31	2.61
(h)	2.96	2.16	2.24	3.82	2.27	2.51
(i)	2.85	2.12	2.20	3.55	2.32	2.45
(j)	3.26	2.22	2.36	4.01	2.30	2.57
(k)	3.16	2.23	2.32	4.19	2.39	2.61
(l)	3.55	2.23	2.41	4.60	2.29	2.69
(m)	2.97	2.21	2.27	3.48	2.41	2.46
(n)	2.02	1.96	1.93	2.74	2.05	2.14

表 A.5: フィンランド語テキスト 14 冊分のべき指数 α。(a)Baskervillen koira, (b)Rikos ja rangaistus, (c)Gorgias, (d)Gulliverin matkat kaukaisilla mailla, (e)Muistelmia kuolleesta talosta, (f)Huckleberry Finnin (Tom Sawyerin toverin) seikkailut, (g)Viidakkopoika, (h)Ylpeys ja ennakkoluulo, (i)Tom Sawyersin seikkailut, (j)Aarresaari, (k)Gil Blas Santillanalaisen elamanvaiheet, (l)Jean-Christophe X Uusi tyopaiva, (m)Kun nukkuja heraa Romaani, (n)Vtanha tarina Montrosesta.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	3.37	2.20	2.37	3.70	2.33	2.49
(b)	3.12	2.23	2.31	3.77	2.37	2.53
(c)	3.32	2.22	2.35	4.19	2.26	2.59
(d)	3.65	2.33	2.47	4.97	2.40	2.39
(e)	3.33	2.19	2.37	4.11	2.35	2.61
(f)	3.17	2.24	2.33	3.93	2.29	2.55
(g)	3.31	2.26	2.34	3.93	2.45	2.60
(h)	3.18	2.12	2.29	3.89	2.12	2.47
(i)	3.35	2.16	2.35	3.89	2.27	2.50
(j)	3.08	2.18	2.29	3.80	2.39	2.51
(k)	3.24	2.25	2.35	4.05	2.36	2.60
(l)	3.48	2.18	2.39	3.86	2.23	2.49
(m)	2.80	2.07	2.16	3.86	2.23	2.49
(n)	2.78	2.11	2.17	3.39	2.26	2.39

表 A.6: ハンガリー語テキスト 14 冊分のべき指数 α。(a)Alomvilag: Elbeszelesek, (b)Elbeszelesek, (c)A lathatatlan ember: Regeny, (d)Szazadunk magyar irodalma kepekben: Szechenyi follepesetol a kiegyezesig, (e)Vezeto elmek: Irodalmi karcolatok, (f)Edes anyafoldem! : Egy nep s egy ember tortenete (1. kotet), (g)Az erkolcsi vilag, (h)Midas kiraly, (i)Torpek es oriasok, (j)A voros regina: regeny, (k)Edes anyafoldem! : Egy nep s egy ember tortenete (2. kotet), (l)Furcsa emberek: Elbeszelesek, (m)Nepmesek Heves- es Jasz-Nagykun-Szolnok-megyebol; Magyar nepkoltesi gyujtemeny 9. kotet, (n)Vegzetes tevedes: Regeny.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	2.31	1.97	1.97	2.65	2.12	2.14
(b)	2.49	2.06	2.06	2.84	2.21	2.22
(c)	2.47	2.07	2.07	2.94	2.19	2.25
(d)	2.57	2.05	2.10	3.18	2.15	2.31
(e)	2.31	1.97	1.97	2.65	2.12	2.14
(f)	2.76	2.18	2.20	3.27	2.23	2.35
(g)	2.73	2.12	2.16	3.39	2.22	2.37
(h)	2.55	1.97	2.05	3.31	2.19	2.35
(i)	2.72	2.14	2.17	3.29	2.21	2.34
(j)	2.32	2.01	2.00	3.20	2.18	2.32
(k)	2.34	1.99	2.00	2.94	2.08	2.21
(l)	2.57	2.00	2.08	3.00	2.07	2.20
(m)	2.24	1.91	1.92	2.81	2.16	2.19
(n)	2.15	1.94	1.91	2.87	2.19	2.22

表 A.7: イタリア語テキスト 14 冊分のべき指数 α。(a)Fra Tommaso Campanella, Vol. 1 la sua congiura, i suoi processi e la sua pazzia, (b)Della storia d'Italia dalle origini fino ai nostri giorni, sommario. v. 1, (c)Ettore Fieramosca: ossia, La disfida di Barletta, (d)In faccia al destino, (e)Fra Tommaso Campanella, Vol. 2 la sua congiura, i suoi processi e la sua pazzia Language, (f)La guerra del Vespro Siciliano vol. 1 Un periodo delle storie Siciliane del secolo XIII, (g)La guerra del Vespro Siciliano vol. 2 Un periodo delle storie Siciliane del secolo XIII, (h)Lettere di Lodovico Ariosto Con prefazione storico-critica, documenti e note, (i)I mesi dell'anno ebraico, (j)Niccolo de' Lapi; ovvero, i Palleschi e i Piagnonia, (k)Le rive della Bormida nel 1794, (l)Novelle, (m)La scienza in cucina e l'arte di mangiar bene Manuale pratico per le famiglie, (n)Orlando Furioso.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	2.63	2.06	2.11	3.05	2.13	2.25
(b)	2.77	2.25	2.24	3.77	2.21	2.50
(c)	2.41	2.09	2.05	2.98	2.20	2.26
(d)	2.39	2.12	2.06	3.10	2.24	2.32
(e)	2.34	1.98	1.99	2.94	2.11	2.22
(f)	2.35	2.00	2.00	2.90	2.09	2.21
(g)	2.56	2.19	2.14	2.95	2.20	2.26
(h)	2.45	2.07	2.05	3.05	2.18	2.28
(i)	2.35	2.03	2.01	3.16	2.14	2.29
(j)	2.41	2.05	2.04	2.99	2.21	2.27
(k)	2.31	2.05	2.01	2.75	2.16	2.17
(l)	2.13	2.01	1.93	2.34	2.15	2.05
(m)	2.43	2.08	2.06	3.22	2.17	2.33
(n)	2.41	2.21	2.09	2.88	2.36	2.28

表 A.8: ポルトガル語テキスト 14 冊分のべき指数 α。(a)Ambicoes: Romance, (b)Amor Crioulo vida argentina, (c)Portugal e Brazil: emigracao e colonisacao, (d)Os deputados brasileiros nas Cortes Geraes de 1821, (e)Os fidalgos da Casa Mourisca Chronica da aldeia, (f)Os Filhos do Padre Anselmo, (g)Historia de Portugal: Tomo I, (h)Viagem ao norte do Brazil feita nos annos 1613 a 1614, pelo Padre Ivo D'Evreux, (i)A Morgadinha dos Cannaviaes (Chronica da aldeia), (j)Obras Completas de Luis de Camoes, Tomo II, (k)A Reforma, (l)Resumo elementar de archeologia christa,(m)Os Trabalhadores do Mar, (n)Tres capitaes.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	2.48	2.11	2.08	2.92	2.16	2.24
(b)	2.36	2.10	2.04	2.82	2.20	2.22
(c)	2.09	1.99	1.90	2.18	2.05	1.95
(d)	2.43	2.12	2.07	2.91	2.16	2.23
(e)	2.19	1.94	1.92	2.27	1.97	1.96
(f)	2.29	1.92	1.95	2.62	2.08	2.10
(g)	2.36	1.96	1.98	2.77	2.05	2.14
(h)	2.32	2.05	2.01	2.92	2.21	2.24
(i)	2.30	2.06	2.01	2.74	2.18	2.18
(j)	2.24	2.06	1.99	3.08	2.23	2.30
(k)	2.08	2.09	2.01	2.82	2.18	2.22
(l)	2.34	2.04	2.01	2.74	2.13	2.17
(m)	2.10	2.01	1.92	3.39	2.21	2.38
(n)	2.48	2.09	2.08	2.90	2.14	2.22

表 A.9: スペイン語テキスト 14 冊分のべき指数 α。(a)El amor, el dandysmo y la intriga, (b)La Argentina La conquista del Rio de La Plata. Poema historico, (c)Cocina moderna, (d)El Criterio, (e)Filosofia Fundamental, (f)Historia de Venezuela, Tomo I, (g)Historia de Venezuela, Tomo II, (h)Las noches mejicanas, (i)Los Merodeadores de Fronteras, (j)Orlando Furioso, Tomo I, (k)Un paseo por Paris, retratos al natural, (l)El Protestantismo comparado con el Catolicismo en sus relaciones con la Civilizacion Europea (Vols 1-2), (m)La Regenta, (n)Su unico hijo.

記号	α_{AX}	α_{AY}	α_{AZ}	α_{BX}	α_{BY}	α_{BZ}
(a)	2.01	1.91	1.85	2.37	2.05	2.02
(b)	2.42	2.02	2.02	2.51	2.06	2.06
(c)	2.13	1.99	1.92	2.40	2.09	2.04
(d)	2.14	1.97	1.92	2.14	1.98	1.92
(e)	2.05	1.94	1.88	2.25	1.99	1.96
(f)	2.13	2.04	1.94	2.34	2.05	2.02
(g)	2.44	2.10	2.07	2.47	2.08	2.08
(h)	2.15	2.00	1.93	2.35	2.06	2.02
(i)	2.12	2.02	1.93	2.33	2.08	2.02
(j)	2.27	2.03	1.98	2.36	2.06	2.01
(k)	2.12	1.97	1.92	2.16	1.99	1.93
(l)	2.07	1.95	1.88	2.17	1.97	1.93
(m)	2.07	1.98	1.90	2.22	2.04	1.97
(n)	2.21	2.04	1.97	2.24	2.04	1.98

表 A.10: 日本語テキスト 14 冊分のべき指数 α。(a) カラマーゾフの兄弟、(b) クリスマスキャロルの夜、(c) フランケンシュタイン、(d) ガリバー旅行記、(e) 宝島、(f) 二都物語、(g) 人間失格、(h) 痴人の愛、(i) 夜明け前　第一部上、(j) 地名の研究、(k) 樋口一葉訳　源氏物語「若菜」、(l) こころ、(m) 三四郎、(n) 坊ちゃん。

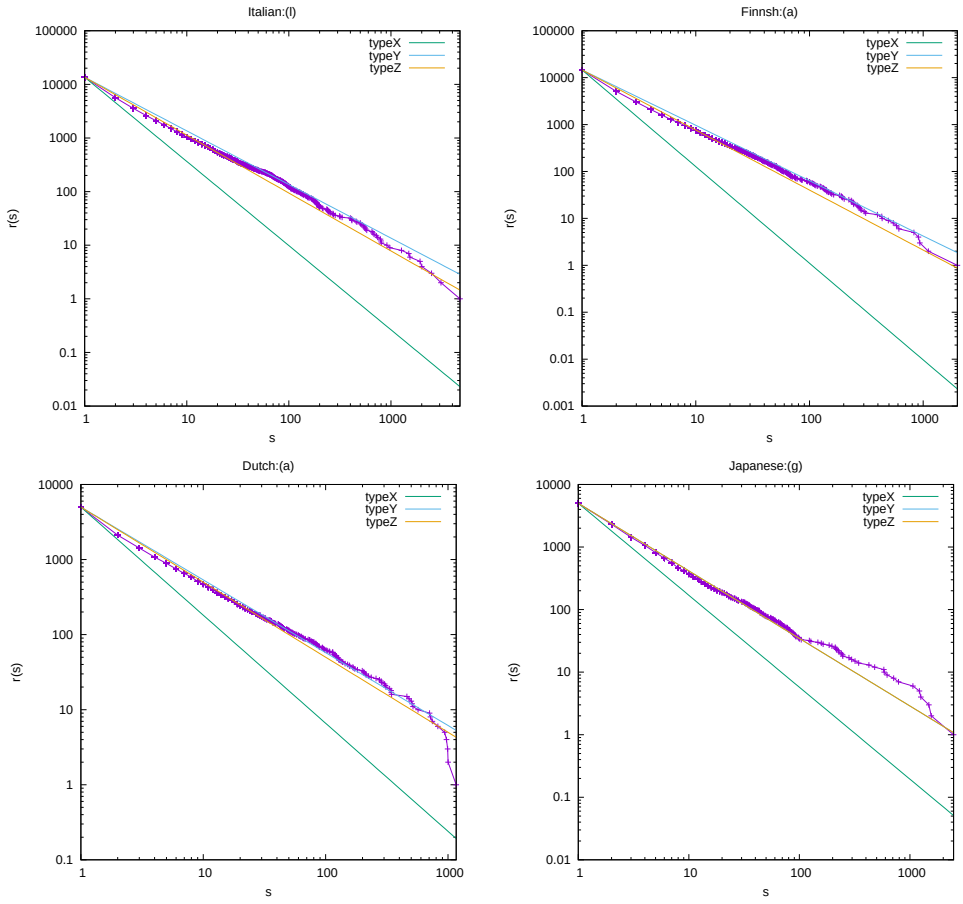

図 A.7: べき指数比較の一例。上列左からイタリア語 (l)Novelle（語彙数制限なし）、フィンランド語 (a)Baskervillen koira（語彙数制限なし）。下列左からオランダ語 (a)Het portret van Dorian Gray（語彙数 5000）、日本語 (g) 人間失格（語彙数 5000）。

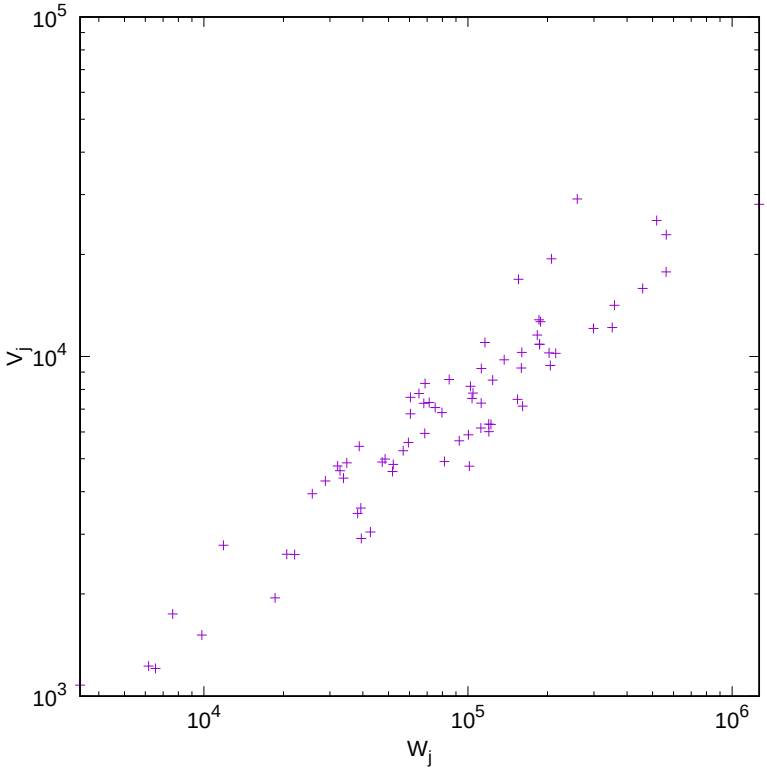

図 A.8: テキスト j の総語彙数 V_j（縦軸）と総単語数 W_j（横軸）の両対数プロット。