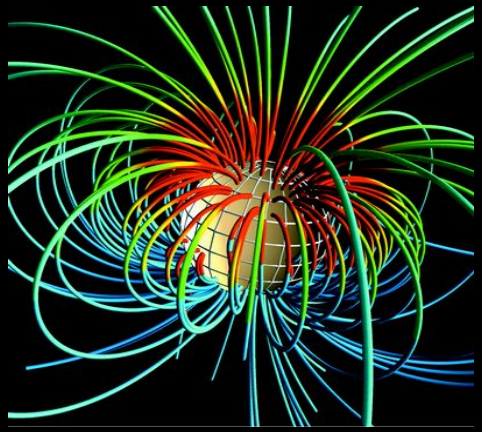
麗澤大学言語研究センターおよび
言語科学会（2008年度会員講習会）共催
2008年7月5日（土） 午後1:00～4:30場所:
麗澤大学生涯教育プラザ1階プラザホール（千葉県柏市）

ワークショップ

日本語コーパスの使用法と解析

エントロピーと冗長度の指標を使ったコーパス共起頻度の分析



玉岡 賀津雄（麗澤大学）
ktamaoka@gc4.so-net.ne.jp

コーパス研究の意味

- 言語のコーパスを，特定の表現の事例を検索するために使うのであれば，事例検索のプログラムを準備するだけでよい．これであれば，出現回数を無視して，表現の種類だけを列挙したのでよいことなる．
- しかし，コーパスでは，特定表現の出現頻度や複数表現の共起頻度を算出することができる．特定の表現について数学的な考察が可能である．

コーパスから得られる情報

- コーパスから得られる情報は、特定の表現が有るか無いかの情報である。
- そのため、「有る」を1とすれば、「無い」は0となる。
- 共起頻度の場合には，表現Aと表現Bの頻度以外に，表現AとBの共起頻度が得られる.
- それぞれについて数えて、足したのが頻度である.

頻度 (frequency)

- 頻度には、重なり頻度(type frequency)と延べ頻度(token frequency)がある。
- 例えば，日本語の母音/a/の日本語の語彙における重なり頻度(Tamaoka & Makioka, 2004)は，6,149,909回である．これは，各単語を1回だけとして，カウントした場合の/a/の出現頻度である．
- 各単語の出現回数を考慮したのが，延べ頻度で，/a/の頻度は124,536,587回である．
- 両頻度にはかなりの違いが見られる。

日本語母音の
重なり頻度と延べ頻度

重なり頻度
(type
frequency)

Table 1 Frequencies of Vowels (φV and +V)

	Vowel					
	/a/	/i/	/u/	/e/	/o/	Total
φV type	6,149,909	43,985,426	39,052,254	4,767,153	14,053,377	108,008,119
Frequency	5.7%	40.7%	36.2%	4.4%	13.0%	100.0%
All Vowel	124,536,587	114,568,843	124,790,724	58,172,645	109,714,325	531,783,124
Frequency	23.4%	21.5%	23.5%	10.9%	20.6%	100.0%

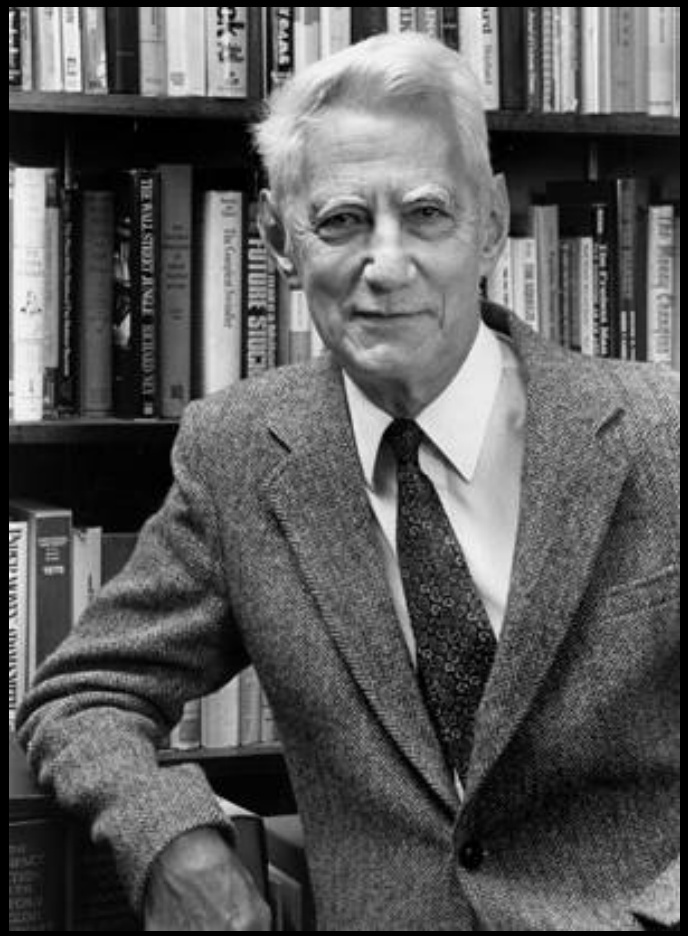
**延べ頻度
(token
frequency)**

コーパス研究のための統計

- 頻度は，メートルで表される長さや分や秒で表される時間と違い，0と1しかないので，従来の統計学のテキストでよく目にする平均や標準偏差などは計算できない．この種の統計解析はノンパラメトリック検定(χ^2分布を利用した一様性および独立性の検定，コレスポンデンス分析，二項ロジスティク回帰，決定木分析など)を使うことが多い．
- しかし，今日は，共起頻度を数学的に指標化する方法について説明する．

クロード・シャノン (Claude Shannon) その1

- アメリカの数学者．マサチューセッツ工科大学教授．情報理論の創始者でデジタル回路の数学的基礎を確立した数学者として知られる．
- Bell Laboratoriesにいた1948年に記念碑的な論文『通信の数学理論(A Mathematical Theory of Communication)』を著し、今日では情報理論の父として知られるようになった．

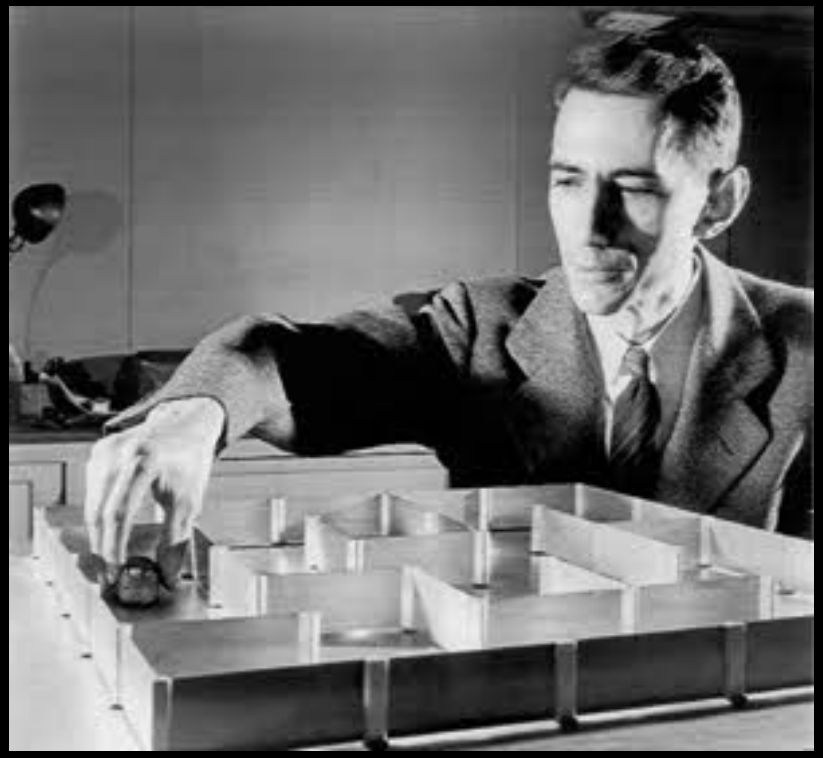


クロード・シャノン (Claude Shannon) その2

- この論文の中で彼は「通信の基本的な問題は、一点にあるメッセージを別のもう一点に正確に、あるいは近似的に再製することにある」との考えを表明し、ある情報を送信するのに1と0の組み合わせを送るだけで十分であることを数学的に示した。これは今日のインターネット、光通信、無線通信などのデジタル通信技術の基盤となっている。

クロード・シャノン (Claude Shannon) その3

- 1916年にミシガン州に生まれた. ミシガン大学で数学と電気工学の学士号を、1940年にはMITから電気工学と数学の博士号を同時に取得して卒業している。
- その後、1941年にBell Laboratoriesに就職、1956年にMITの客員教授、1958年には教授になる。



クロード・シャノン (Claude Shannon) その4

- 修士論文「A Symbolic Analysis of Relay and Switching Circuits」はデジタル回路の数学的基盤が1と0で全てを記述できるブール代数にあることを示したもので、これによりシャノンは近代的なスイッチング理論の創始者ともなった。戦時中にシャノンは暗号に関する研究に携わった.
- 1949年「Communication Theory of Secrecy Systems」を著し、単なる複雑なパズルのように思われていた暗号の分野を科学の域に高め、暗号学の基礎を確立した。
- 2001年2月26日，84歳で死去.

共起頻度パターンの指標化

クロード・シャノン(Claude Shannon) (1948)

『通信の数学理論』
(A Mathematical Theory of Communication)

エントロピー(entropy)
冗長度(redundancy)

- シャノンの『通信の数学理論』で，特に有名な概念が，「エントロピー(entropy)」と「冗長度(redundancy)」である.
- もともと，0と1からなる情報理論から出発したものであり，同じように0と1で表せるコーパス研究の単語などの使用頻度にも応用できる.
- 本ワークショップでもシャノンのエントロピーと冗長度を，コーパスから得られた共起頻度の解析に応用してみる.

エントロピー (entropy: H)

$$H = -\sum_{j=1}^{J} p_j \log_2 p_j$$

エントロピー (entropy: H)

- 情報量の尺度の一つ．平均情報量H(entropy)は以下の公式で定義する．

$$H=-\sum_{j=1}^{j} p_j \cdot \log_2 p_j$$

- あいまいさや乱雑度の増減を示す指標

(有本, 1982; 堀, 1979; 海保, 1989; 玉岡・宮岡・林, 2003; Tamaoka, Lim & Sakai, 2004を参照)

エントロピーの最大値 (H_{max})

- エントロピーが最大であるとは，すべてが等しい確率で生起する場合である。
- いずれが起こっても不思議ではない混沌とした無秩序の状態

- 重なり頻度(表現の数)をJとすると，

$$H_{max}=\log_2 J$$

によりエントロピー最大値が得られる。

冗長度 (redundancy: R)

- 表現の多様性と使用頻度から一つの値を算出して，無駄の程度を表すことができる。

$$R=(1-H/H_{max})\times 100\ (\%)$$

H: エントロピー

H_{max}: エントロピーの最大値

100倍してパーセントで表す

エントロピーと冗長度の指標を組み合わせることで，ある共起表現の多様性と規則性を，絶対的な出現頻度に左右されることなく簡単な数値で表すことができる．

注: とはいえ，ある程度の出現頻度がなくては，エントロピーの計算はできないので，大規模コーパスを使用しなくてはならない．

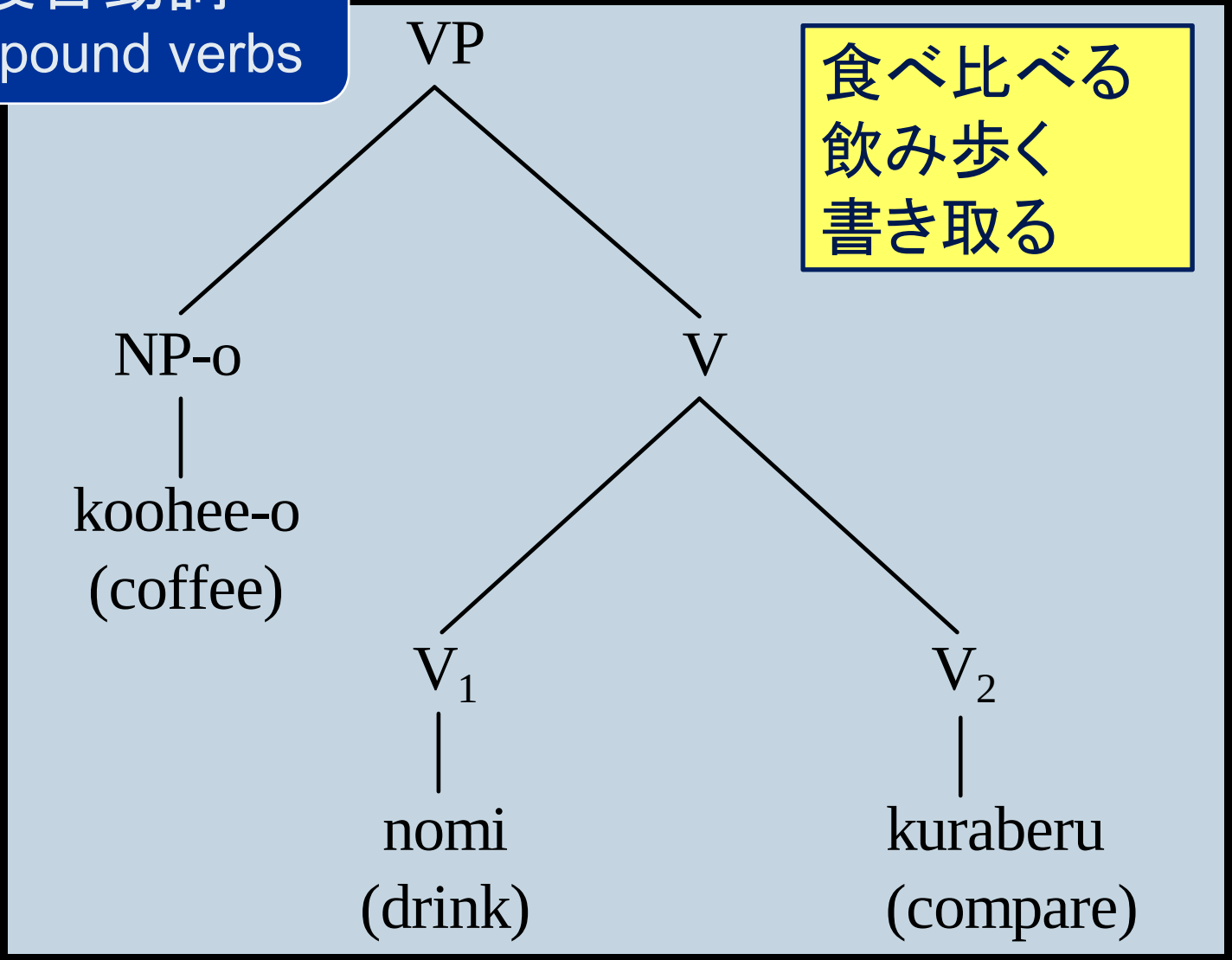
Tamaoka, K., Lim, H., & Sakai, H. (2004). Entropy and redundancy of Japanese lexical and syntactic compound verbs. *Journal of Quantitative Linguistics*, *11(3)*, 233-250.

- 複合動詞に使われる2つの動詞の結合頻度をエントロピーと冗長度で調べてみる.

日本語の動詞ー複合動詞

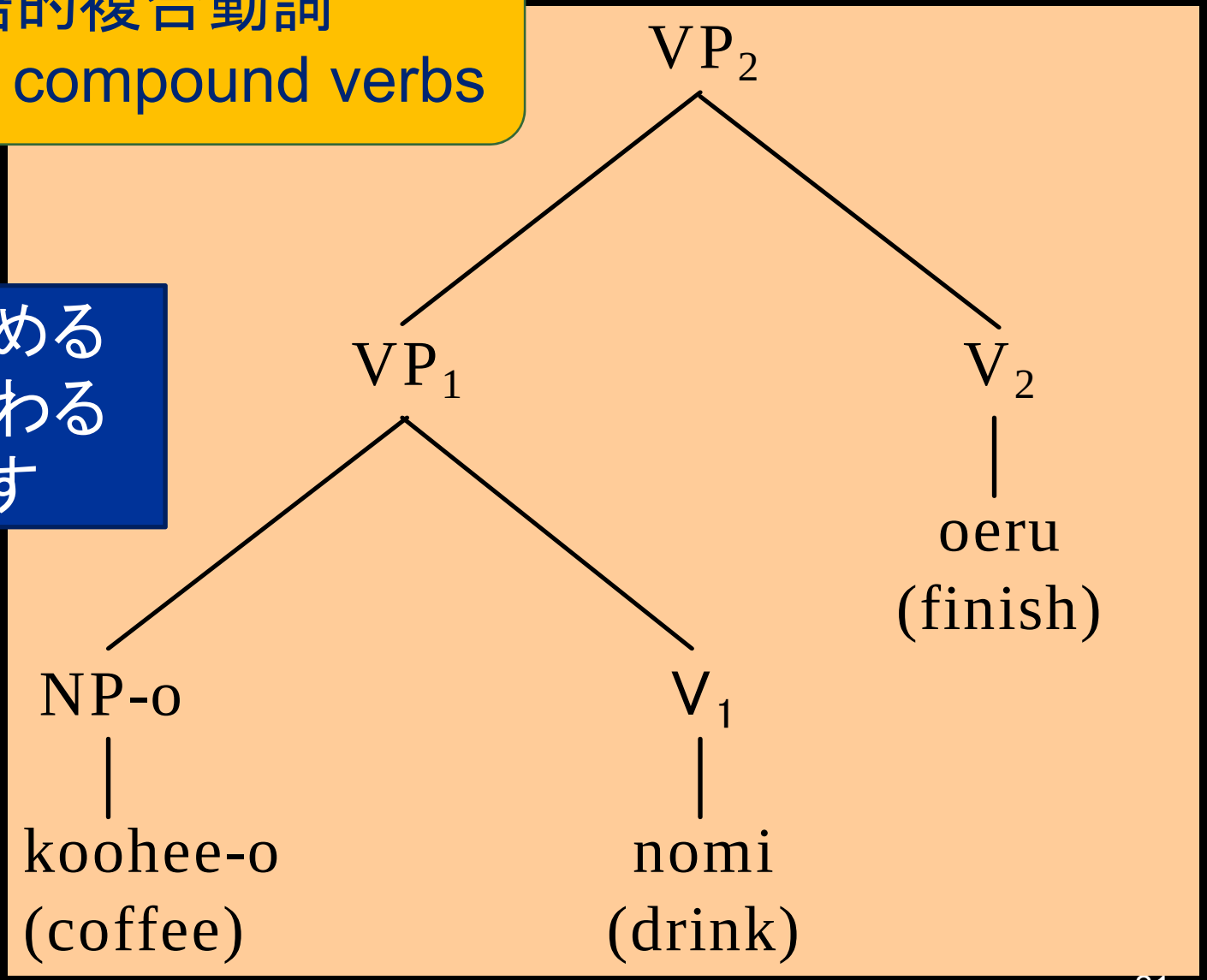
- 日本語では，2つの動詞が組み合わされることが多い．これらを複合動詞(compound verbs)という．
- これらの複合動詞は，語彙的複合動詞(lexical compound verbs)と統語的複合動詞(syntactic compound verbs)の2種類がある．(影山, 1993, 1999a, 1999b).

語彙的複合動詞
Lexical compound verbs



統語的複合動詞
Syntactic compound verbs

食べ始める
飲み終わる
書き直す



研究の仮説

語彙的複合動詞と統語的複合動詞の比較

- 語彙的複合動詞と統語的複合動詞では，2つの動詞の共起パターンが異なると思われる.
- 語彙的複合動詞は，2つの動詞が一定の組み合わせで出現すると予想されるので，冗長度が高く，エントロピーは低いのではないかと思われる.
- 一方，統語的複合動詞は，二つ目の動詞(V2)が多様な一つ目の動詞(V1)と結合すると考えられるので，エントロピーが高く，冗長度は低くなると思われる.
- 語彙的と統語的複合動詞の両方のV2になる動詞は検索から外した.

コーパスの検索装置

- アメリカのパデュー大学先端技術言語学習研究所の深田淳が作成した日本語用例・コロケーション抽出システム『茶漉』を使用した。
  http://tell.fll.purdue.edu/chakoshi/index2.html
- 『茶漉』はコーパスから用例およびコロケーション情報を抽出するシステムである。
- このシステムは、「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（代表者：当時，名古屋大学国際言語文化研究科日本言語文化専攻教授・大曽美恵子）科学研究費補助金によるプロジェクトの一環として開発されたものである。

『茶漉』の名前の由来

- 『茶漉(ちゃこし)』という名称の由来は，コーパスを検索可能なデータファイルに変換する段階で形態素解析システム『茶筌(ちゃせん)』（奈良先端科学技術大学院大学自然言語処理学講座開発による）を用いるが，茶筌を用いて立てたお茶（データ）から必要な情報のみを漉(こ)し取って取り出すシステムということで『茶漉』という。

『茶漉』の小説コーパス
青空文庫コーパス

- 青空文庫コーパス

  青空文庫 (http://www.aozora.gr.jp)に収録されている文学作品のうち、現代語で書かれているものを選んでコーパス化したもの。
- 『茶漉』で検索できる青空文庫コーパスの総語数は8,370,720語。
- 作品例:『地図に出てくる男女』吉行エイスケ,『ごん狐』新美南吉など

『茶漉』の新聞コーパス
毎日新聞

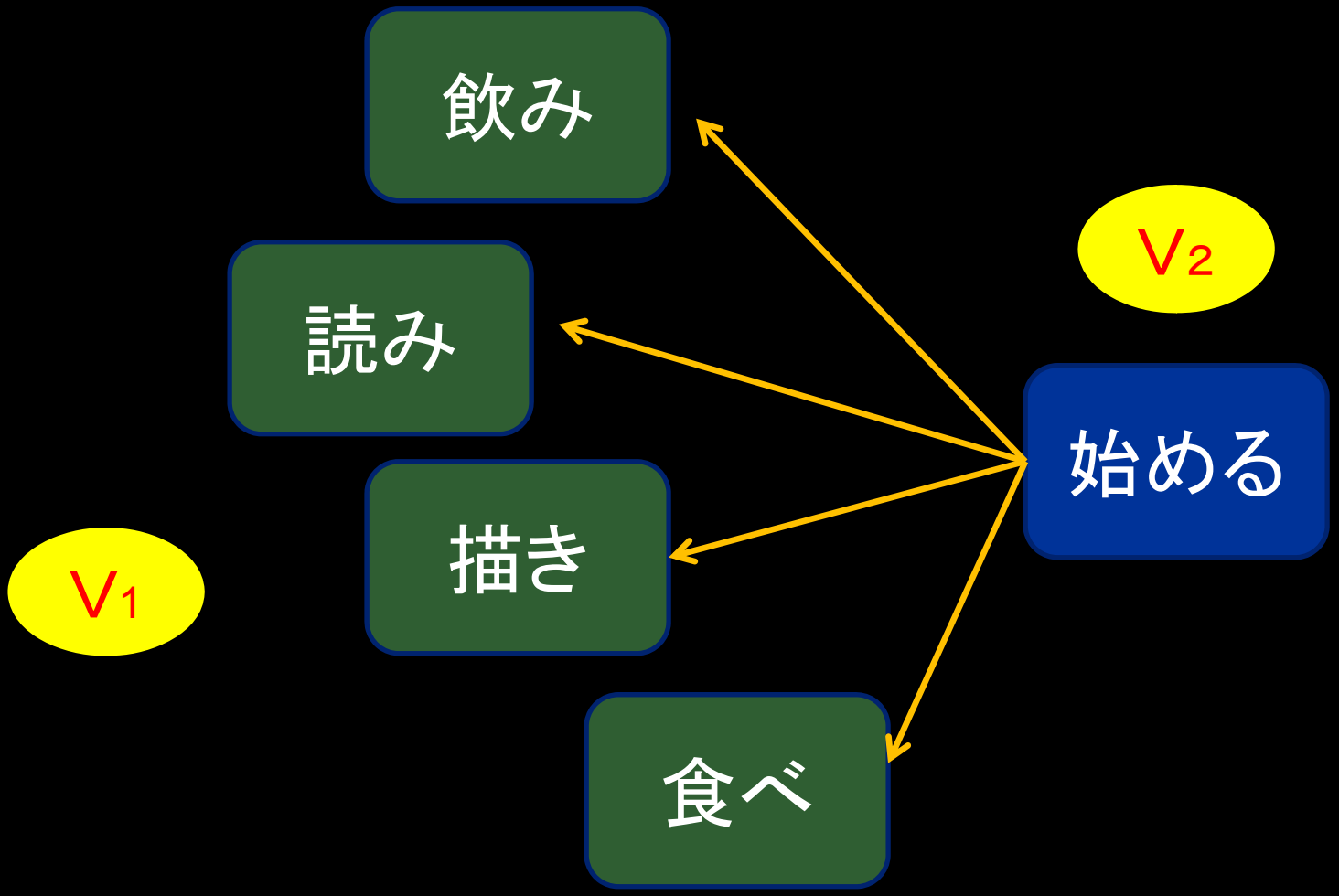
- 毎日新聞（1991年～1999年）
- 毎日新聞の全記事を１年ごとにファイルにまとめたもの。
- 現在，９年分が『茶漉』で使用可能。
- 総語数は，273,514,662語。

複合動詞の研究で使ったコーパス
『茶漉』から2種類のコーパスを使用

- 本研究では，毎日新聞と青空文庫の2種類のコーパスを使用した.
- 毎日新聞は，1991年から1994年までの4年間の記事で，延べ頻度は88,454,573語である.
- 48種類の動詞(V_2)を88の語彙的複合動詞と21種類の統語的複合動詞から選んだ.
- 青空文庫は，明治から昭和のはじめに書かれた小説で，現代語で書かれたものだけを選んでいる．総語彙数は，8,370,720語である.

毎日新聞のコーパス4年分から語彙的および統語的複合動詞を検索する.

複合動詞V_1とV_2の(例えば,「飲み始める」,「食べ始める」)検索には,V_2(例えば,「始める」)を固定して,V_1に来る動詞(例えば,「飲み」,「食べ」)を検索する.その際に,延べ頻度が10以下の場合は,そのV_2は分析に採用しないことにした.


飲み
読み
描き
食べ
始める
V2
V1

V_1 とV_2「始める(*hajimeru*)」の複合動詞

飲み 始める	飲み	のみ	1188	9
読み 始める	読み	よみ	2050	9
描き 始める	描き	えがき	950	8
語り 始める	語り	かたり	3326	8
考え はじめる	考え	かんがえ	23958	8
向け 始める	向け	むけ	20075	7
やり 始める	やり	やり	2690	7
書き 始める	書き	かき	2973	6

毎日新聞のコーパスでは，207種類の
V_2「始める」と結合する統語的複合動詞があった.

統語的複合動詞「抜く」の毎日新聞4年分での検索例

ID	複合語の用例	V1動詞	V1頻度	V1+V2頻度
1	愛し 抜く	愛し	83	1
2	歩き 抜く	歩き	384	3
3	い ぬく	い、射	77,576	4
4	いじめ 抜く	いじめ	672	1
5	選び 抜く	選び	954	2
6	踊り 抜く	踊り	409	2
7	がんばり 抜く	り	418	5
8	嫌い ぬく	嫌い	439	1
9	苦しみ ぬく	苦しみ	158	1
10	し 抜く	し	632,236	7

複動動詞は23種類で131回の
V_1とV_2共起頻度(述べ頻度)

エントロピー　　$H=-\Sigma p_j \log_2 p_j$

p_j ・・・「抜く」V_2のV_1との特定表現の共起頻度全体での比率

$\mathrm{Log}_2 p_j$・・・比率を2を底とする対数で表現した数値

$p_j \log_2 p_j$ ・・・比率とその対数値を掛けた値

$\Sigma p_j \log_2 p_j$ ・・・それらを積算した値

$-\Sigma p_j \log_2 p_j$ ・・・－1を掛けた値

これがエントロピー(H)である.

ID	複合語の用例	V1動詞	V1頻度	V1+V2頻度	Rate	Log(数値, 2)	掛け算
1	愛し 抜く	愛し	83	1			
2	歩き 抜く	歩き	384	3			
3	い ぬく	い、射	77,576	4			
4	いじめ 抜く	いじめ	672	1			
5	選び 抜く	選び	954	2			
6	踊り 抜く	踊り	409	2			
7	がんばり 抜く	り	418	5			

Rate ・・・p_j　全体の共起頻度が131なので,
　1÷131が入る.

対数・・・$\mathrm{Log}_2 p_j$　Rateを対数にした値. +log(pj ,2)

Rateと対数値を掛けた値.　$p_j \log_2 p_j$

やってみよう！

- Excelのファイル「2008.7.5 - 「抜く」複合動詞の共起頻度」を読み込んで，一緒にエントロピー，エントロピー最大値，冗長度を計算してみよう.

統語的複合動詞の エントロピーと冗長度の例

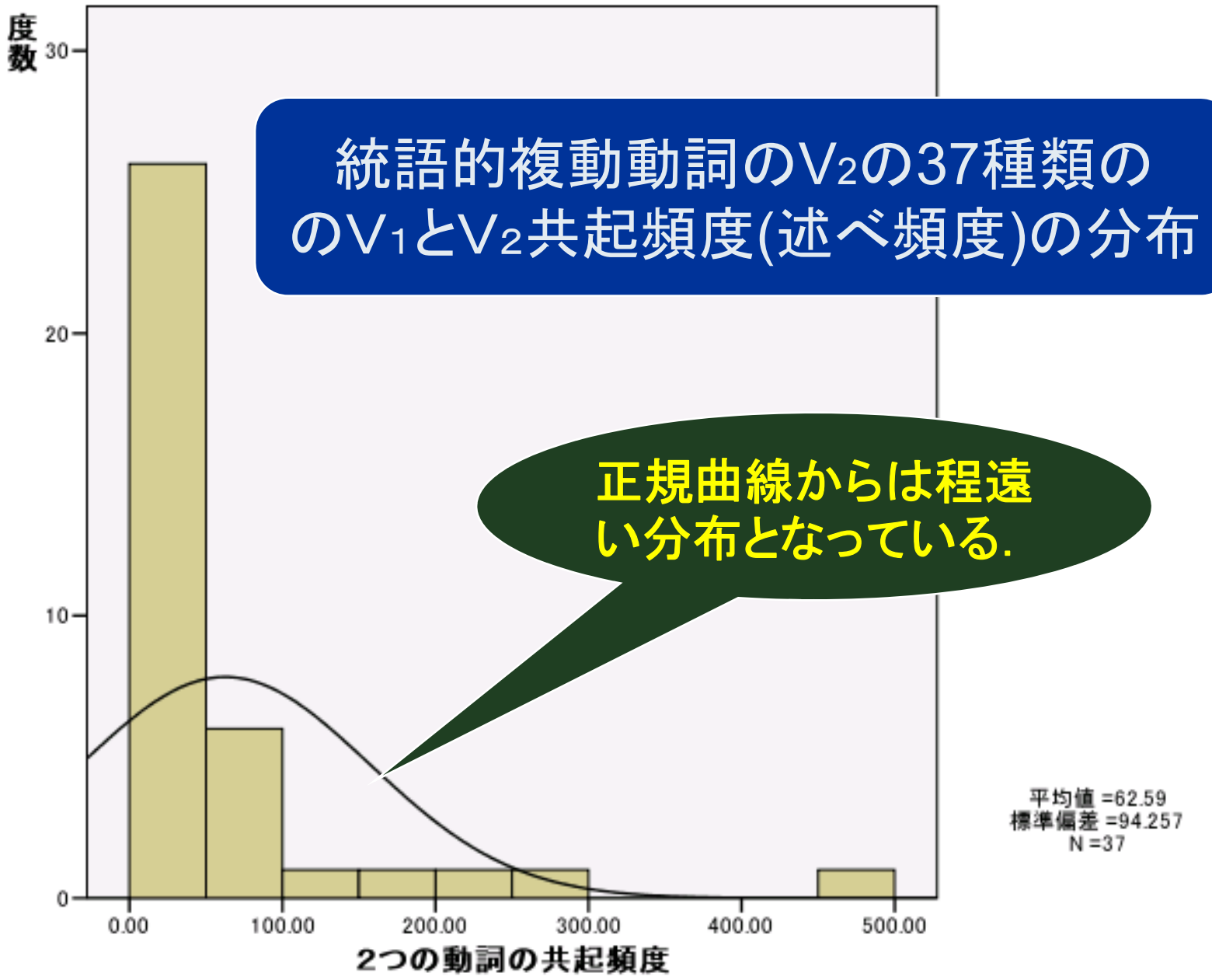
V_2 verbs		V_2 token frequency	V_1 total token	V_1 type frequency	V_1&V_2 token	Entropy	Redundancy (%)
Japanese	Phonetic						
続ける	tuzukeru	5,519	539,169	261	1425	6.73	16.21
始める	hazimeru	2,983	1,379,861	207	657	6.50	15.55
あう	au	2,302	295,787	170	873	6.16	16.87
過ぎる	sugiru	3,777	368,408	130	515	5.71	18.74
まくる	makuru	86	708,256	32	66	4.56	8.91
終わる	owaru	1,884	51,545	31	56	4.50	9.18
終える	oeru	503	850,402	24	37	4.31	5.90
尽くす	tukusu	687	843,270	26	89	3.72	20.86
ぬく	nuku	575	724,584	23	131	3.11	31.33
かねる	kaneru	328	1,062,433	18	108	2.82	32.27

Tamaoka, Lim and Sakai (2004)のPDFファイル

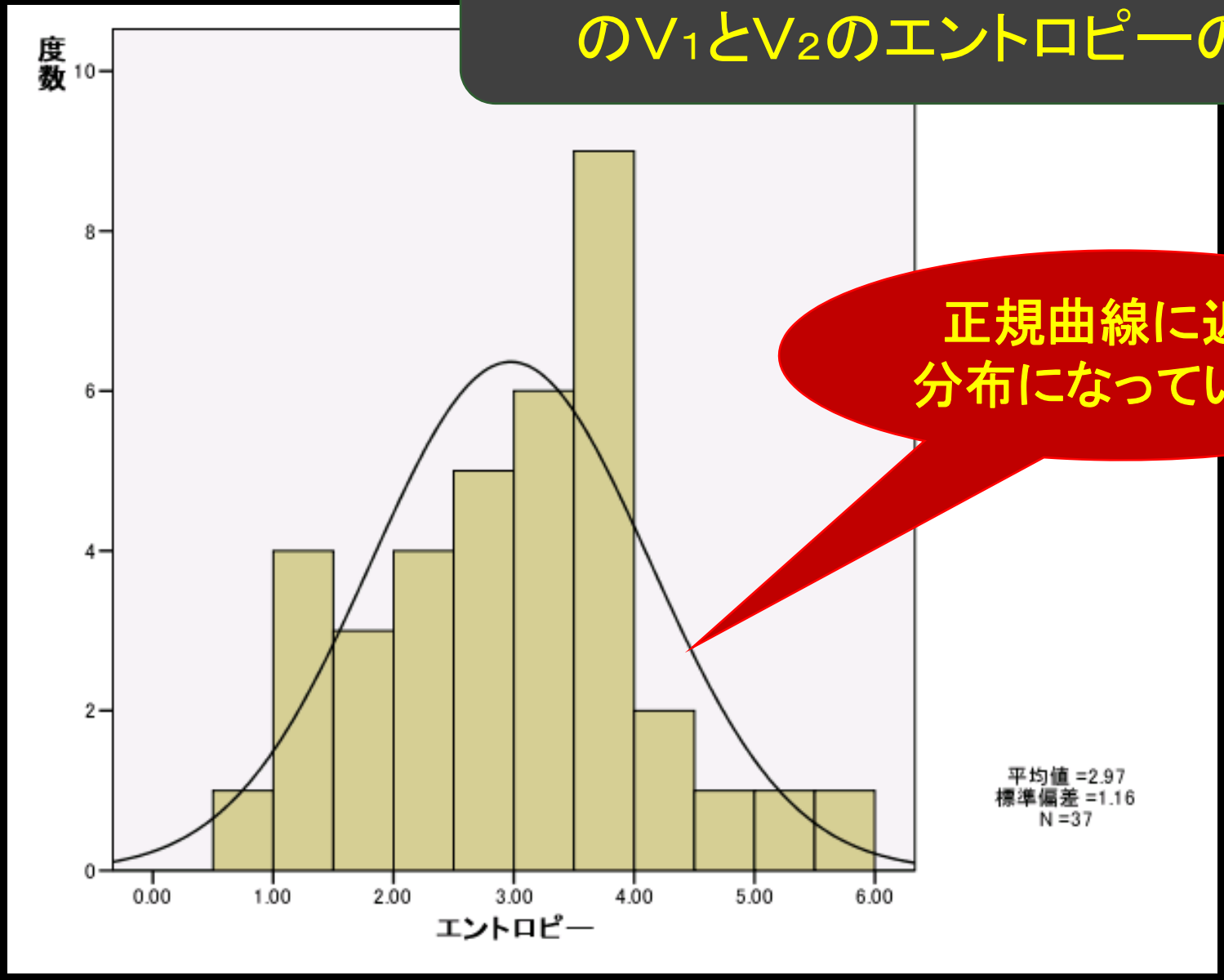
語彙的複合動詞の
エントロピーと冗長度の例

V_2 verbs		V_2 token frequency	V_1 total token	V_1 type frequency	V_1&V_2 token	Entropy	Redundancy (%)
Japanese	Phonetic						
込む	komu	295	1,098,690	81	278	5.76	9.10
あげる	ageru	2,914	45,880	57	174	5.30	9.20
切れる	kireru	543	64,292	44	119	4.66	14.73
取る	toru	5,947	53,493	33	94	4.39	13.04
回る	mawaru	1,021	17,989	27	61	4.27	10.12
つく	tuku	2,354	8,906	19	45	3.81	10.34
歩く	aruku	1,554	30,414	18	44	3.78	9.35
上がる	agaru	1,808	40,283	31	229	3.69	25.56
継ぐ	tugu	355	20,382	15	33	3.68	5.88

語彙的複合動詞37種類の共起頻度，エントロピー，冗長度の分布をみてみよう.

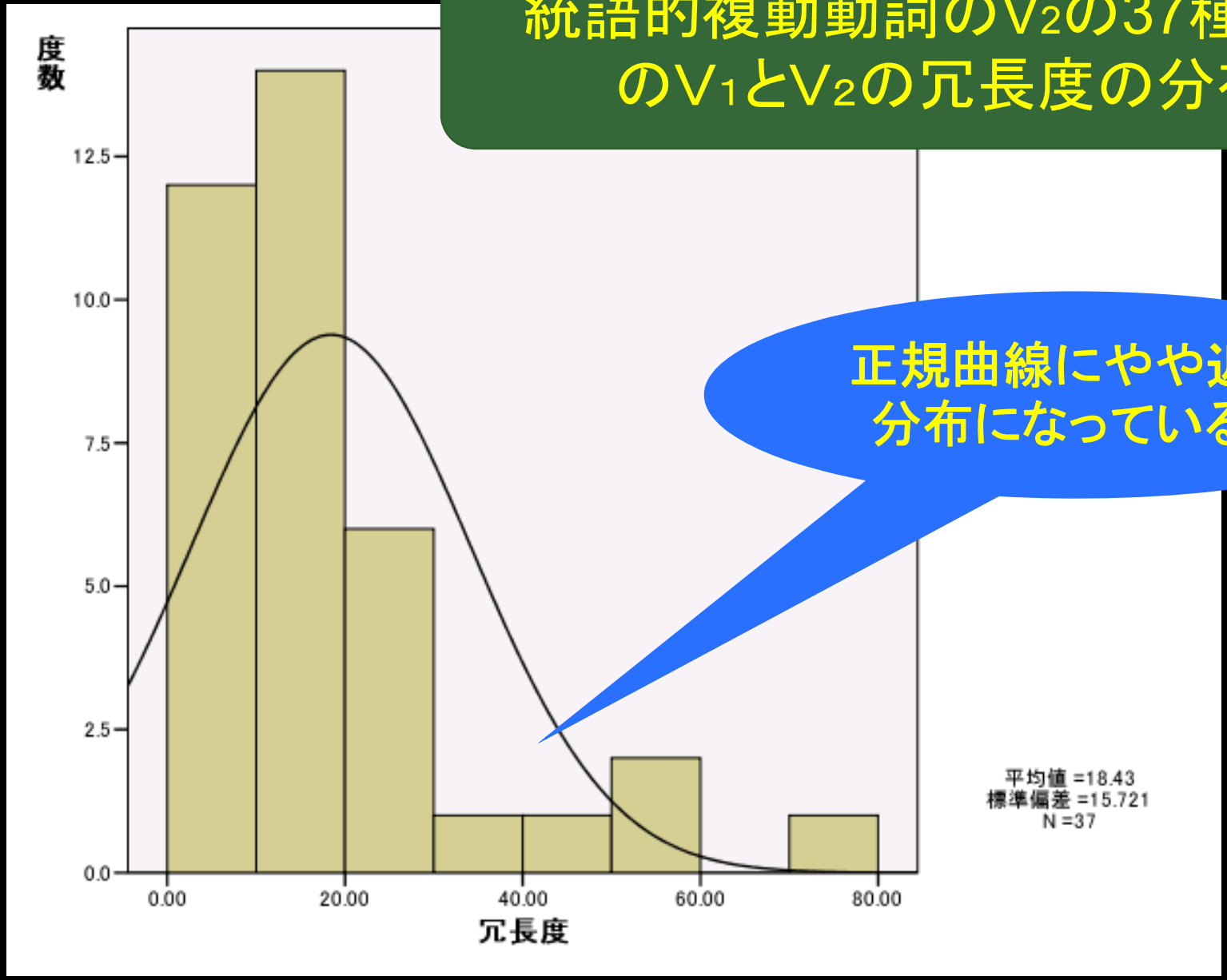

統語的複動動詞のV2の37種類の
のV1とV2共起頻度(述べ頻度)の分布
正規曲線からは程遠い分布となっている.
度数
30
20
10
0
0.00
100.00
200.00
300.00
400.00
500.00
2つの動詞の共起頻度
平均値 =62.59
標準偏差 =94.257
N =37

統語的複動動詞のV2の37種類のV1とV2のエントロピーの分布



正規曲線に近い
分布になっている.

統語的複動動詞のV2の37種類のV1とV2の冗長度の分布



正規曲線にやや近い分布になっている.

パラメトリック分析

- 言語コーパスから得られた共起頻度や頻度は，正規分布とは程遠い分布を示す．そのため，正規分布を要求するパラメトリック分析はできないことになる．ノンパラメトリック分析(カイ二乗分布を使った統計手法など)
- 一方，エントロピーと冗長度は，ほぼ正規分布，あるいはそれに近い分布を示すので，パラメトリック分析が可能である．

すべての複合動詞の計算後

- エントロピーと冗長度の計算が終われば，次に，語彙的複合動詞と統語的複合動詞をパラメトリック統計で分析することができます．
- 統計的な処理にはSPSSを利用しますが，これは今回の講習会では，時間がないので詳細は説明しません．

SPSSを使って分析

一元配置の分散分析 (one-way analysis of variance; ANOVA)

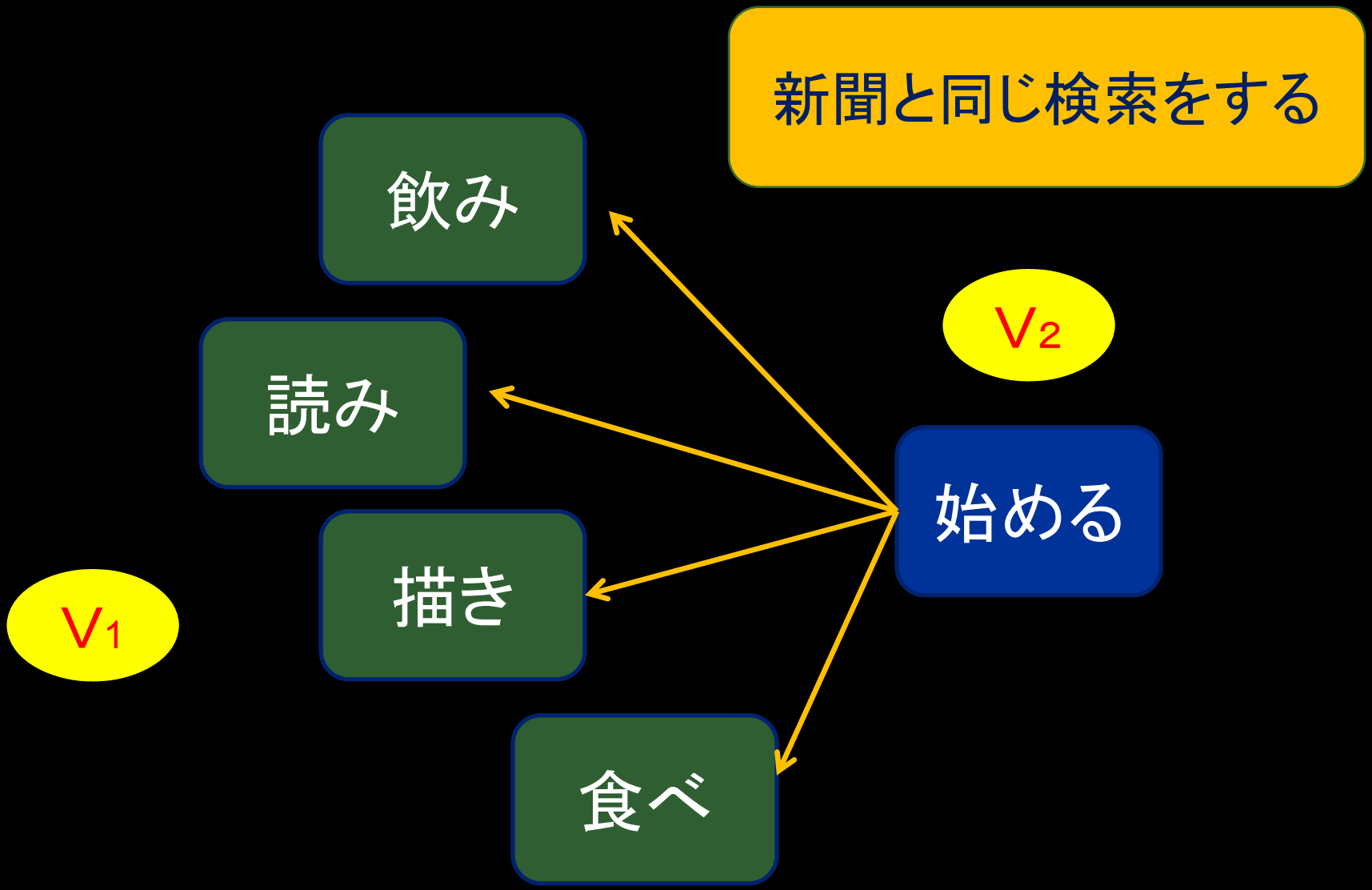
- 毎日新聞のコーパスから得た統語的複合動詞と語彙的複合動詞のV_2の共起頻度パターンを分散分析で分析する. (独立したサンプルの *t*検定でもできる)
- エントロピーについて一元配置の分散分析の結果，語彙的複合動詞(*n*=37, M=2.97, SD=1.16)の方が，統語的複合動詞(*n*=11, M=4.38, SD=1.95)よりもエントロピーが有意に低かった[$F(1,46)=8.95$, $p<.01$].
- 冗長度については，語彙的複合動詞と統語的に違いはなかった[$F(1,46)=0.31$, *n.s.*].

結果の解釈

- 二つの動詞(V_1とV_2)が組み合わされて作られる複合動詞について，語彙的複合動詞の方が，統語的複合動詞よりも二つの動詞の結びつきが規則的であり，２つの動詞が一つの単位として規則的に(idiosyncratic)結合していることが分かる.
- 一方，統語的複合動詞は，V_2の動詞が多様なV_1の動詞と結びつくため，多様な2つの動詞V_1とV_2の組み合わせを作っているようである.
- 冗長度には違いがないので，特定のV_1とV_2の結合が頻繁に繰り返されるといったことはないようである.

小説ー青空文庫コーパス

- 青空文庫コーパス(http://www.aozora.gr.jp)
- 『茶漉』で検索できる青空文庫コーパスの総語数は8,370,720語.
- 『茶漉』には，青空文庫 に収録されている文学作品のうち、現代語で書かれているものだけを選んでコーパス化している.
- 作品例:『地図に出てくる男女』吉行エイスケ，『ごん狐』新美南吉など.


新聞と同じ検索をする
飲み
読み
描き
食べ
始める
V2
V1

一元配置の分散分析

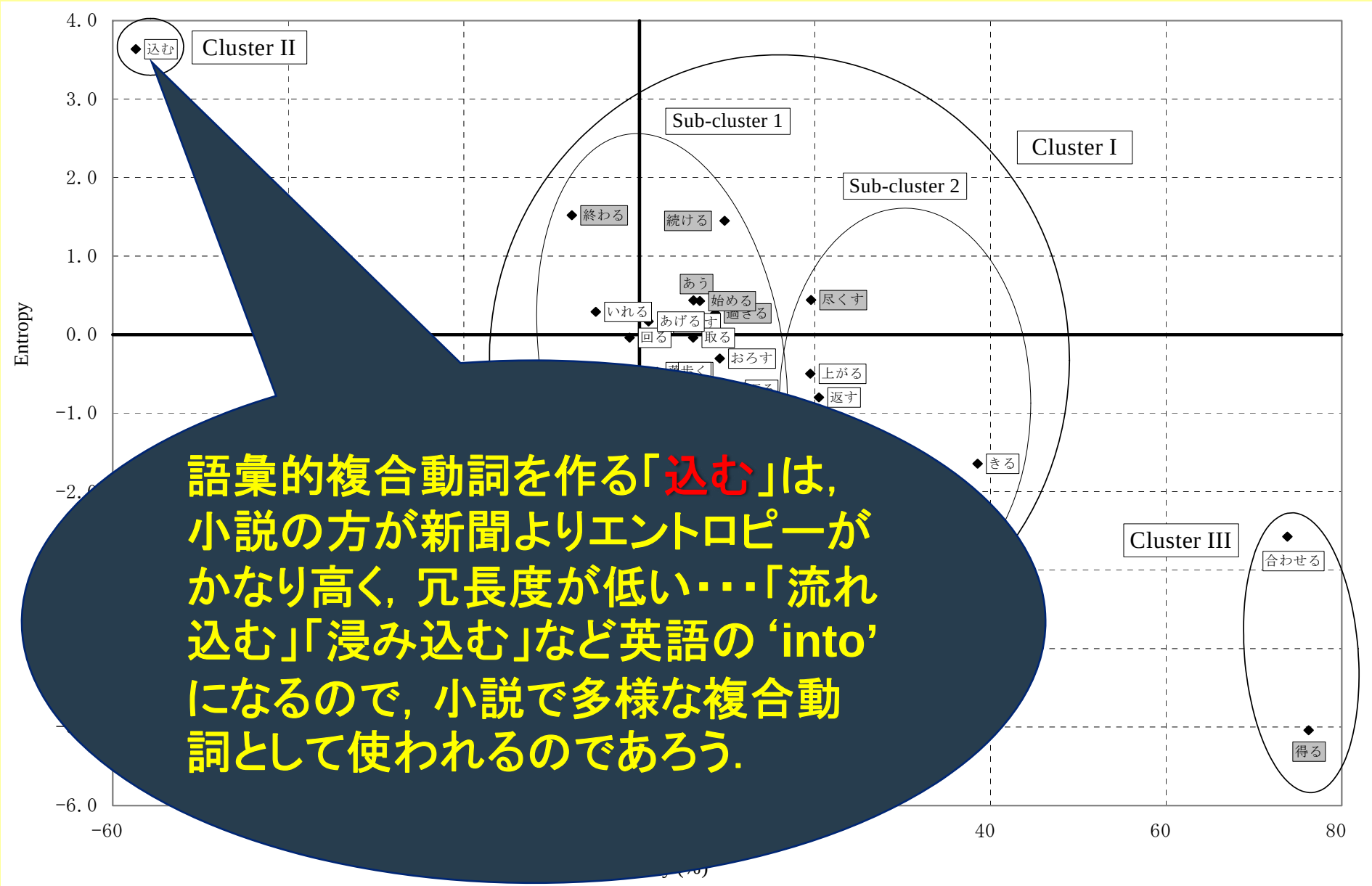
- 青空文庫のコーパスから得た統語的複合動詞と語彙的複合動詞のV_2の共起頻度パターンを分散分析で分析する.
- エントロピーについて一元配置の分散分析の結果，語彙的複合動詞(n=29, M=3.72, SD=0.90) の方が，統語的複合動詞(n=8, M=4.86, SD=1.12)よりもエントロピーが有意に低かった[$F(1,35)=9.14$, $p<.01$].
- 冗長度については，語彙的複合動詞と統語的に違いはなかった[$F(1,35)=0.01$, *n.s.*].

結果の解釈

青空文庫のコーパスも，毎日新聞のコーパスの分析結果と同じであった.

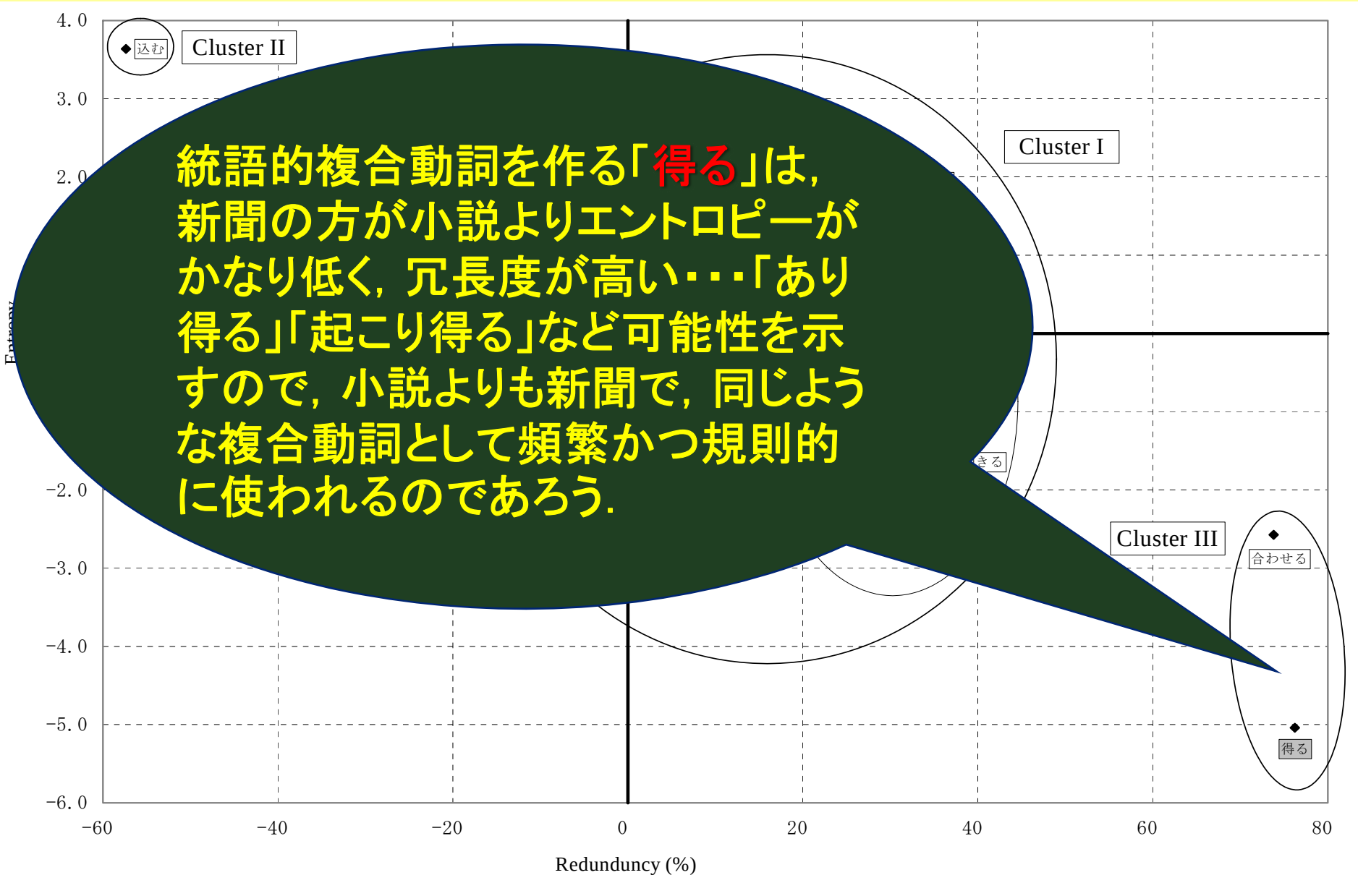
新聞と小説での複合動詞の共起頻度の違いの比較

- 新聞と小説での複合動詞の2つの動詞の共起頻度パターンを比較するために，小説から新聞のエントロピーと冗長度を引いて，それを二次元のグラフに描いた.
- 0に近いほど両者に違いがないことになる.



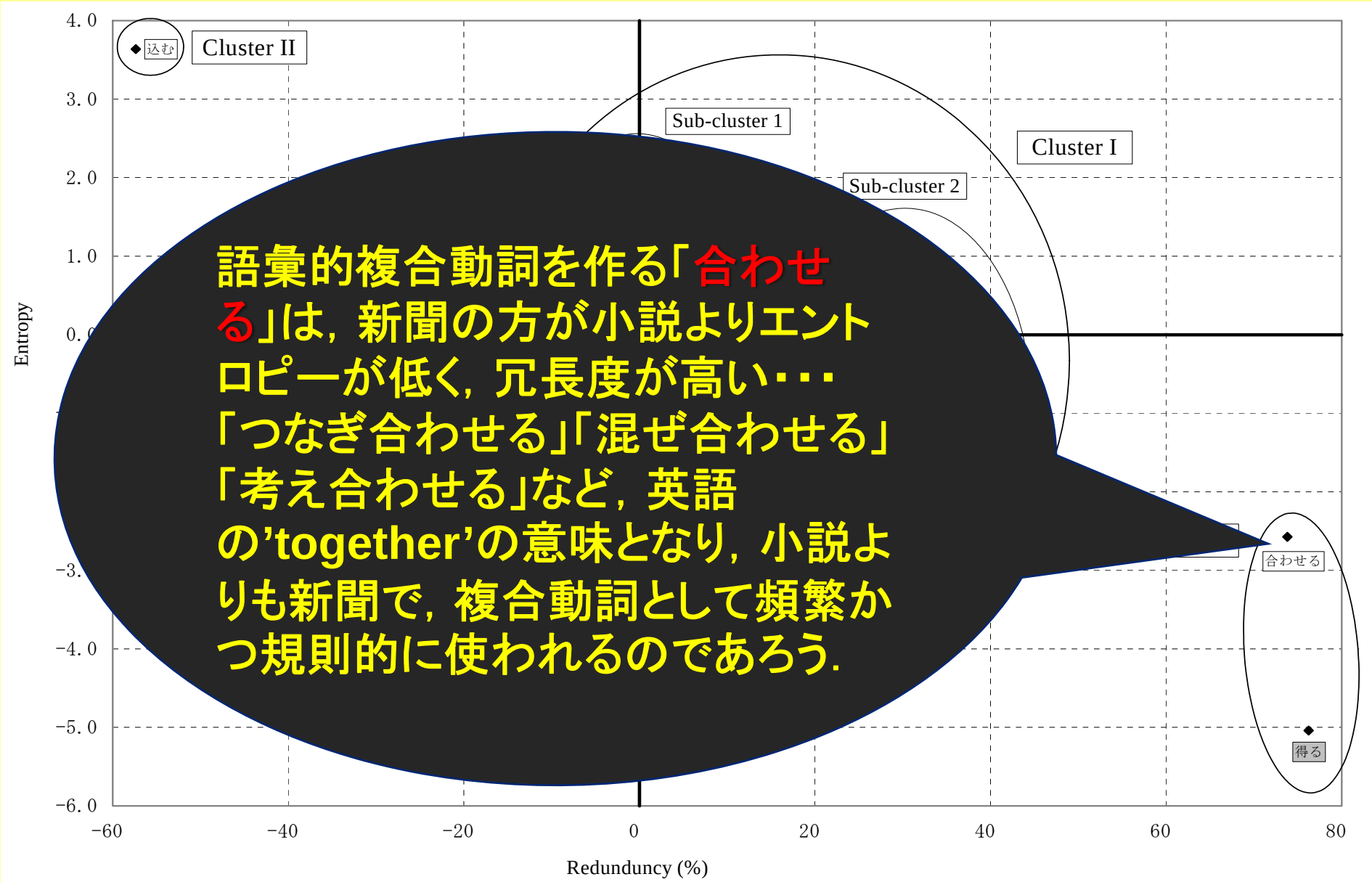
re 4. Plotting and cluster of lexical and syntactic compound V_2 verbs based on differences of entropy and redundancy between the corpus of the newspaper and the no

Note 1 : Hierarchical cluster analysis using Ward's method with the square Euclidean distance formed three clusters including two sub-clusters.
Note 2 : Compound V_2 verbs (n =34) in this figure were selected from overlapped items between Figure 1 and Figure 2.
Note 3 : Verbs in shadowed boxes are lexical compound verbs (n =26) while verbs in unshadowed boxes are syntactic compound verbs (n =8).
Note 4 : Differences were calculated from entropy and redundancy of the newspaper corpus subtracted from those of the novel corpus.



re 4. Plotting and cluster of lexical and syntactic compound V_2 verbs based on differences of entropy and redundancy between the corpus of the newspaper and the no

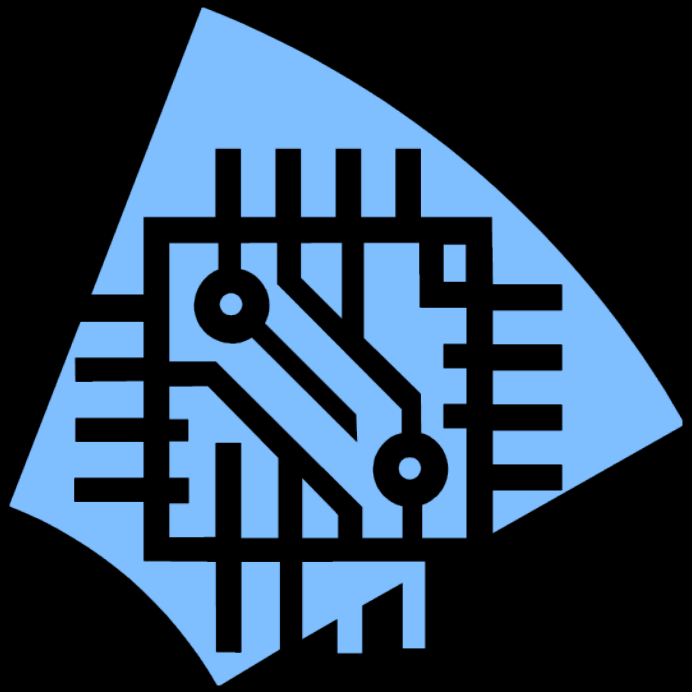
Note 1 : Hierarchical cluster analysis using Ward's method with the square Euclidean distance formed three clusters including two sub-clusters.
Note 2 : Compound V_2 verbs (n =34) in this figure were selected from overlapped items between Figure 1 and Figure 2.
Note 3 : Verbs in shadowed boxes are lexical compound verbs (n =26) while verbs in unshadowed boxes are syntactic compound verbs (n =8).
Note 4 : Differences were calculated from entropy and redundancy of the newspaper corpus subtracted from those of the novel corpus.



re 4. Plotting and cluster of lexical and syntactic compound V_2 verbs based on differences of entropy and redundancy between the corpus of the newspaper and the no

Note 1 : Hierarchical cluster analysis using Ward's method with the square Euclidean distance formed three clusters including two sub-clusters.
Note 2 : Compound V_2 verbs (n =34) in this figure were selected from overlapped items between Figure 1 and Figure 2.
Note 3 : Verbs in shadowed boxes are lexical compound verbs (n =26) while verbs in unshadowed boxes are syntactic compound verbs (n =8).
Note 4 : Differences were calculated from entropy and redundancy of the newspaper corpus subtracted from those of the novel corpus.

玉岡賀津雄・木山幸子・宮岡弥生(2008). ヒトの言語産出とコーパスの頻度はどのくらい類似しているか，*日本言語学会第136回大会予稿集*(学習院大学)，122-127..



- オノマトペと動詞に共起頻度をエントロピーと冗長度で調べてみる.

新聞と小説のコーパスの違い

新聞: 多くの新聞記者によって広く情報を伝達するために書かれる。

小説: ある特定の作家によって書かれた独特の文体や表現を持つ。

両資料は，書き手と目的において大きく異なる。

コーパス研究の目的と本研究の位置づけ

ヒトの言語産出について一般化した規則を見いだすことをコーパス研究の目的とするならば，種々のコーパスがヒトの産出とどの程度類似しているかを検討する必要がある。

オノマトペと動詞の共起の頻度パターンを新聞と小説のコーパスとヒトの産出で比較

オノマトペを選択する際の条件

① 誰でも知っていると思われる基本的なオノマトペであること。
② ひらがなで表記される畳語で，オノマトペであることが誰の目にも明らかであること。
③ 様態の副詞として使えること。
④ 形容動詞としては使えないこと。
⑤ 「する」以外の動詞とも一緒に用いるのが一般的であるもの。

本研究で選んだ28種類のオノマトペ

どんどん，だらだら，がんがん，ばたばた，
ばりばり，すらすら，ゆらゆら，ちょろちょろ，
ことこと，ぽたぽた，ぱちぱち，ころころ，
しとしと，かんかん，とぼとぼ，ぐらぐら，
ぴょんぴょん，きらきら，ぼうぼう，すやすや，
ぐうぐう，めそめそ，ぷんぷん，じろじろ，
ごくごく，しくしく，ずきずき，げらげら

コーパスの検索装置

- 深田淳先生が作成した日本語用例･コロケーション抽出システム『茶漉』を使用

http://tell.fll.purdue.edu/chakoshi/index2.html

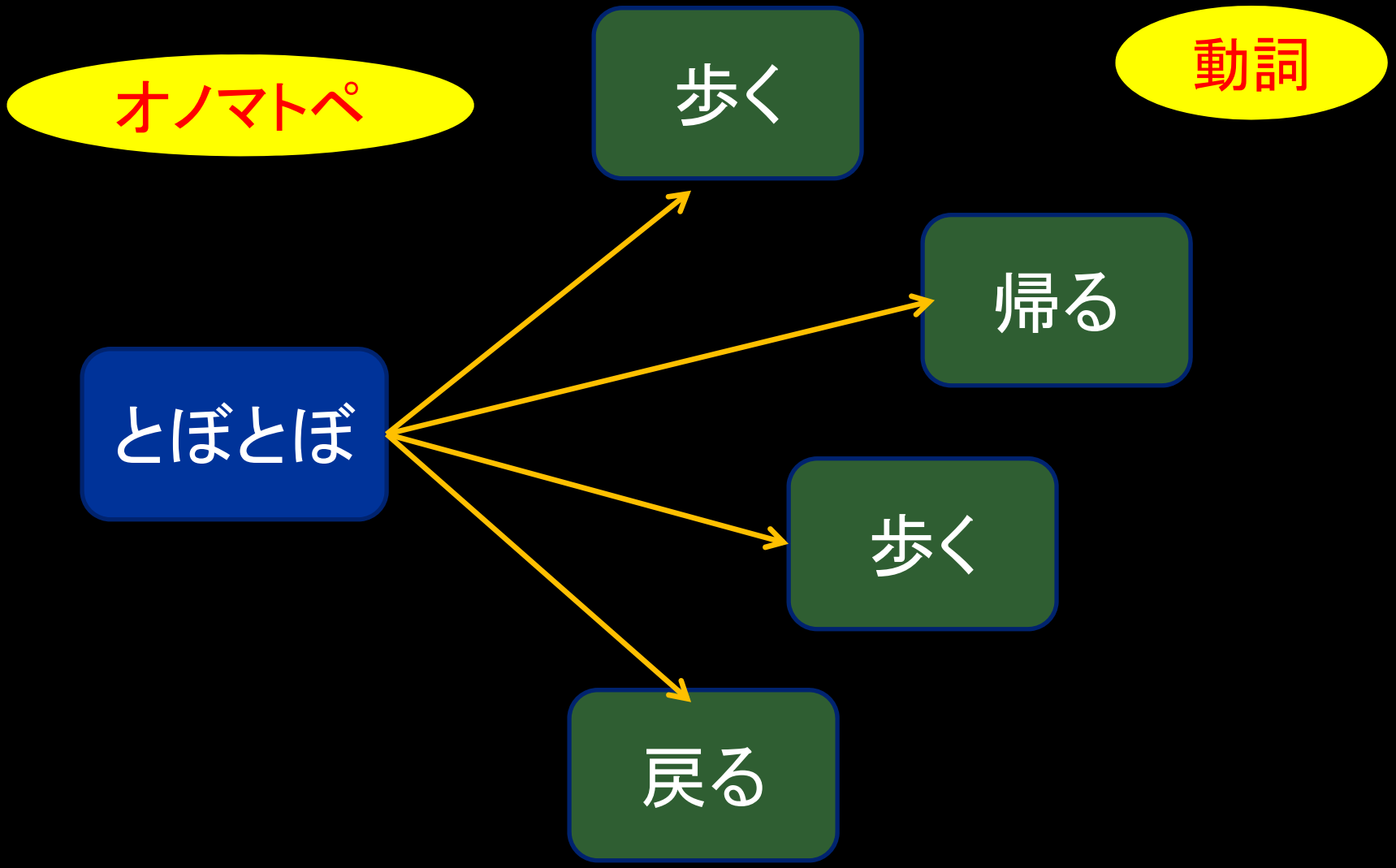
小説ー青空文庫コーパス

- 青空文庫コーパス

  青空文庫 (http://www.aozora.gr.jp)に収録されている文学作品のうち、現代語で書かれているものを選んでコーパス化したもの。
- 『茶漉』で検索できる青空文庫コーパスの総語数は8,370,720語。
- 作品例:『地図に出てくる男女』吉行エイスケ,『ごん狐』新美南吉など

新聞ー毎日新聞

- 毎日新聞（1991年～1999年）
- 毎日新聞の全記事を1年ごとにファイルにまとめたもの。
- 現在，9年分が『茶漉』で使用可能。
- 総語数は，273,514,662語。


オノマトペ
動詞
歩く
帰る
とぼとぼ
歩く
戻る

例文抽出後のオノマトペと動詞の共起頻度カウントのための基準

1. オノマトペと動詞が共起していない文は、分析から除外した。
   - 述語が省略されている文
   - 述語があっても動詞ではない文（形容詞・名詞述語等）

2. オノマトペと共起する動詞の分類基準

ヒトーオノマトペから動詞を産出

- 36名の大学生

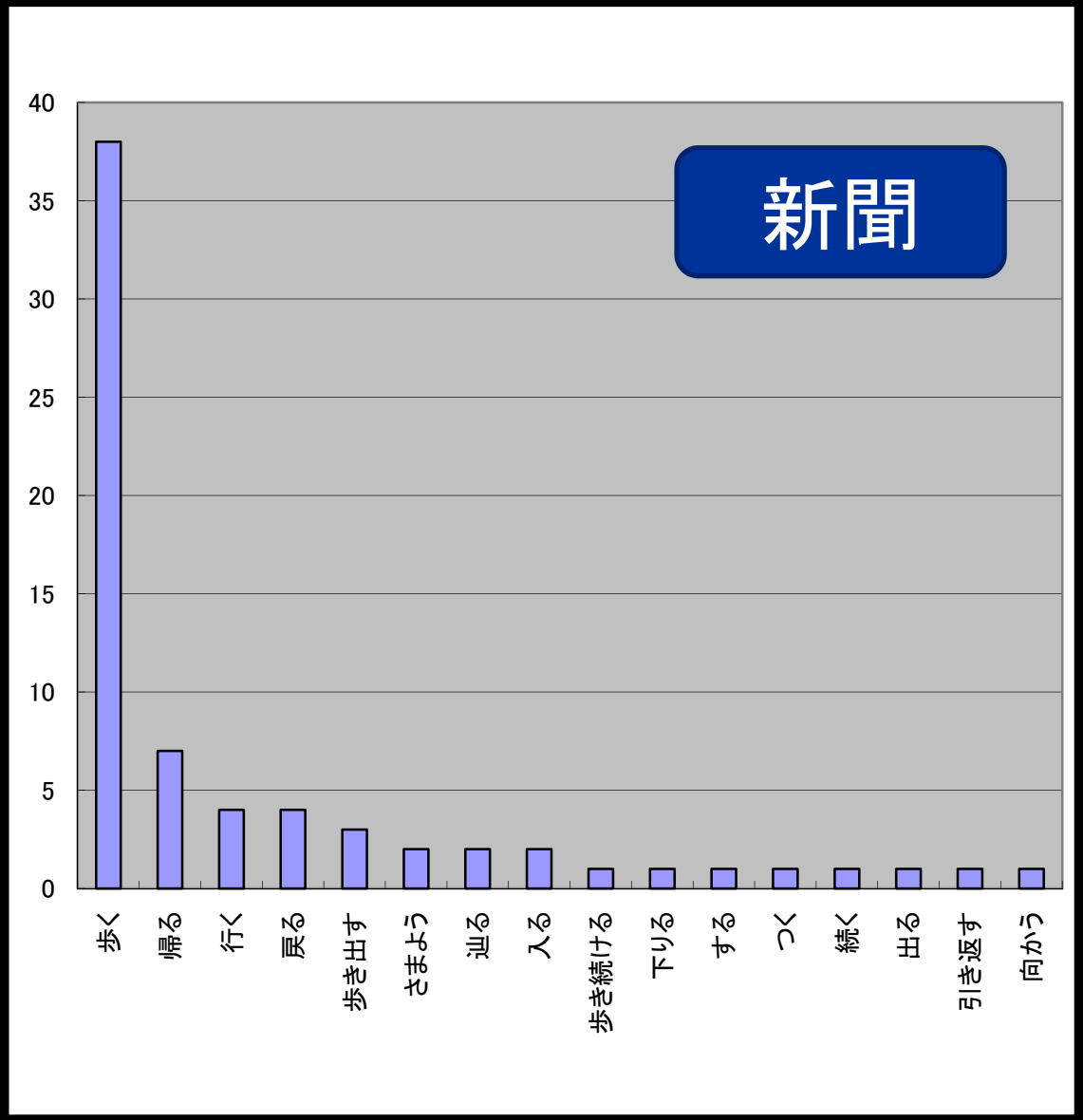
  (最年少18歳7カ月，最年長21歳7カ月; 平均20歳4カ月，標準偏差1歳2カ月; 男性32名，女性4名)
- オノマトペと共起すると思われる動詞を30秒で思いつく限り挙げてもらった。
- オノマトペと動詞の共起頻度を算出した。

エントロピーと冗長度

エントロピーと冗長度の指標を組み合わせることで，ある表現の多様性と規則性を，絶対的な出現頻度に左右されることなく簡単な数値で表すことができる。

新聞での共起頻度

	とぼとぼ	
1	歩く	38
2	帰る	7
3	行く	4
4	戻る	4
5	歩き出す	3
6	さまよう	2
7	辿る	2
8	入る	2
9	歩き続ける	1
10	下りる	1
11	する	1
12	つく	1
13	続く	1
14	出る	1
15	引き返す	1
16	向かう	1
	合　計	70

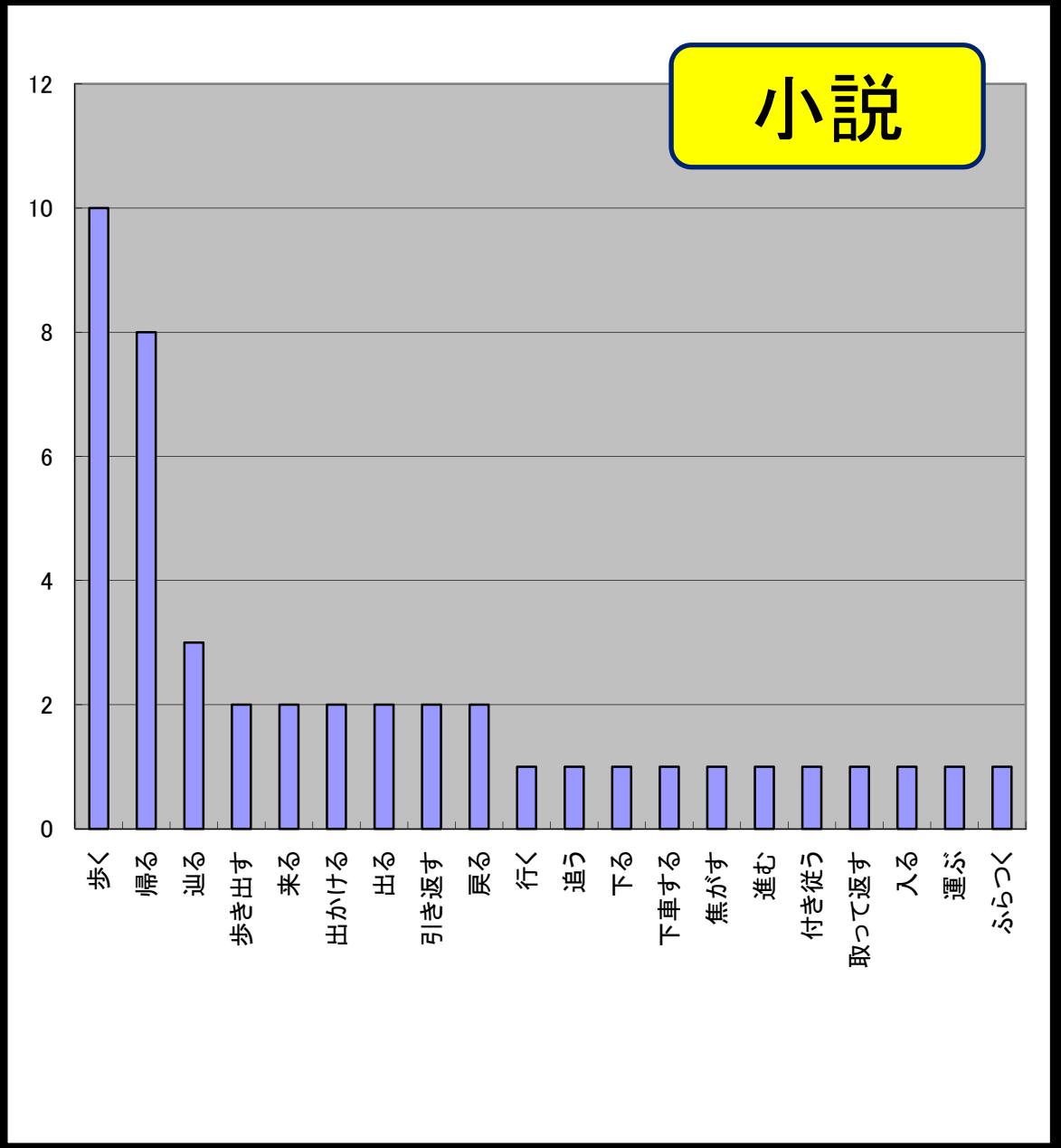


「とぼとぼ」と動詞の新聞コーパスから得た共起頻度からエントロピーと冗長度を計算してみよう！

- Excelのファイル「2008.7.5 – 新聞「とぼとぼ」動詞の共起頻度」を読み込んで，一緒にエントロピー，エントロピー最大値，冗長度を計算してみよう.

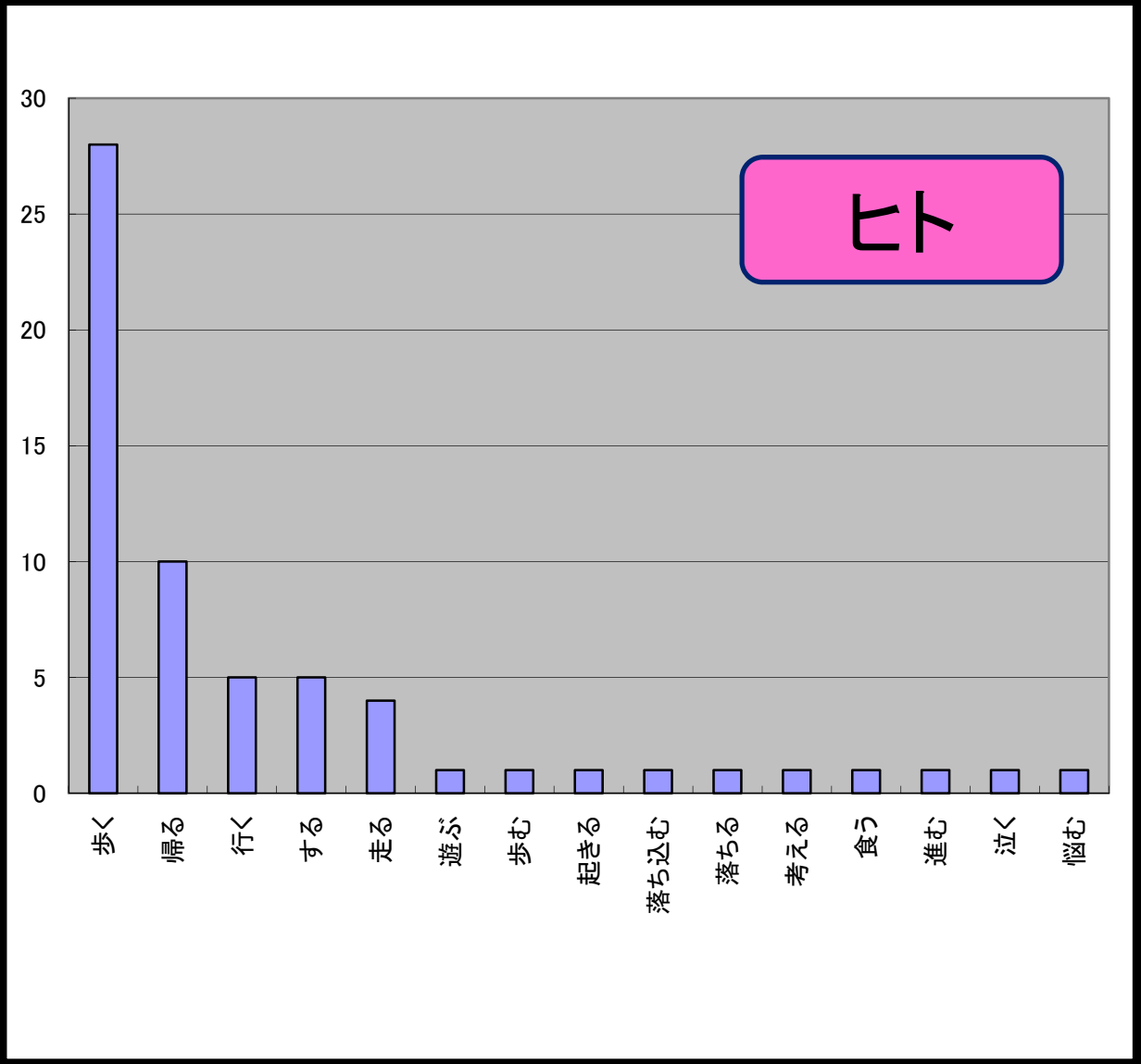
小説での共起頻度

	とぼとぼ	
1	歩く	10
2	帰る	8
3	辿る	3
4	歩き出す	2
5	来る	2
6	出かける	2
7	出る	2
8	引き返す	2
9	戻る	2
10	行く	1
11	追う	1
12	下る	1
13	下車する	1
14	焦がす	1
15	進む	1
16	付き従う	1
17	取って返す	1
18	入る	1
19	運ぶ	1
20	ふらつく	1
	合　計	44



ヒトの産出での共起頻度

	とぼとぼ	
1	歩く	28
2	帰る	10
3	行く	5
4	する	5
5	走る	4
6	遊ぶ	1
7	歩む	1
8	起きる	1
9	落ち込む	1
10	落ちる	1
11	考える	1
12	食う	1
13	進む	1
14	泣く	1
15	悩む	1
	合計	62



練習1ー小説

「とぼとぼ」と動詞の小説コーパスから得た共起頻度からエントロピーと冗長度を計算してみよう！

- Excelのファイル「2008.7.5 – 小説「とぼとぼ」動詞の共起頻度」を読み込んで，一緒にエントロピー，エントロピー最大値，冗長度を計算してみよう.

練習2ーヒト

「とぼとぼ」と動詞のヒトの産出から得た共起頻度からエントロピーと冗長度を計算してみよう！

- Excelのファイル「2008.7.5 – ヒト「とぼとぼ」動詞の共起頻度」を読み込んで，一緒にエントロピー，エントロピー最大値，冗長度を計算してみよう.

表1 ヒトのオノマトペに対する動詞の産出と新聞および小説のコーパスによる共起頻度

オノマトペ	ヒトの動詞の産出				新聞のコーパスにおける動詞の共起				小説のコーパスにおける動詞の共起			
	動詞種類	産出頻度	エントロピー	冗長度	動詞種類	共起頻度	エントロピー	冗長度	動詞種類	共起頻度	エントロピー	冗長度
どんどん	48	118	4.851	13.14%	926	4424	7.808	20.78%	132	241	6.451	8.43%
だらだら	40	112	4.604	13.49%	71	220	4.723	23.20%	19	28	3.982	6.27%
がんがん	35	83	4.523	11.81%	90	234	5.603	13.69%	14	26	3.377	11.30%
ばたばた	27	77	4.095	13.88%	89	317	4.522	30.17%	66	108	5.627	6.91%
ばりばり	28	80	4.045	15.86%	40	152	4.076	23.41%	20	28	4.155	3.86%
すらすら	27	110	3.750	21.13%	28	62	4.222	12.17%	35	55	4.663	9.08%
ゆらゆら	26	99	3.640	22.56%	51	134	4.490	20.84%	34	49	4.844	4.79%
ちょろちょろ	19	81	3.549	16.45%	20	34	3.925	9.17%	19	26	4.056	4.52%
ことこと	18	52	3.374	19.10%	7	22	2.154	23.29%	7	8	2.750	2.04%
ぱたぱた	18	78	3.246	22.15%	7	19	1.878	33.09%	19	36	3.857	9.20%
ぱちぱち	15	65	3.244	16.98%	27	60	4.042	15.00%	20	79	2.903	32.84%
ころころ	13	63	3.122	15.63%	32	179	3.358	32.83%	21	35	3.971	9.59%
しとしと	14	51	3.005	21.07%	3	12	1.041	34.33%	6	8	2.500	3.29%
かんかん	12	49	2.902	19.06%	11	40	2.504	27.61%	22	43	4.108	7.87%
とぼとぼ	15	62	2.744	29.77%	16	70	2.617	34.56%	20	44	3.778	12.58%
ぐらぐら	13	61	2.657	28.20%	20	87	2.867	33.66%	18	46	3.027	27.42%
ぴょんぴょん	14	91	2.473	35.06%	8	50	3.367	19.25%	9	13	3.085	2.68%
きらきら	13	88	2.300	37.84%	42	442	2.582	52.12%	32	99	3.472	30.56%
ぼうぼう	8	43	2.238	25.41%	5	7	2.128	8.35%	7	9	2.725	2.92%
すやすや	10	59	2.211	33.43%	9	39	2.095	33.89%	15	48	2.935	24.88%
ぐうぐう	9	43	2.158	31.93%	5	13	1.506	35.15%	12	29	3.147	12.21%
めそめそ	13	82	2.096	43.36%	4	23	1.445	27.76%	7	33	2.171	22.66%
ぶんぶん	8	49	2.051	31.65%	9	41	1.694	46.57%	10	27	2.088	37.14%
じろじろ	10	61	2.040	38.59%	8	56	1.306	56.48%	21	96	3.162	28.01%
ごくごく	12	52	1.964	45.21%	4	16	1.186	40.69%	6	10	2.161	16.40%
しくしく	9	42	1.901	40.02%	8	16	2.781	7.31%	10	55	2.510	24.45%
ずきずき	9	68	1.791	43.51%	5	10	2.046	11.86%	4	7	1.664	16.78%
げらげら	6	50	1.346	47.91%	8	31	1.964	34.55%	6	16	2.046	20.84%
平均	17.46	70.32	2.93	26.94%	55.46	243.21	3.00	27.21%	21.82	46.50	3.40	14.27%
標準偏差	10.23	21.44	0.93	10.98%	169.32	811.14	1.52	12.65%	24.60	46.29	1.09	10.26%

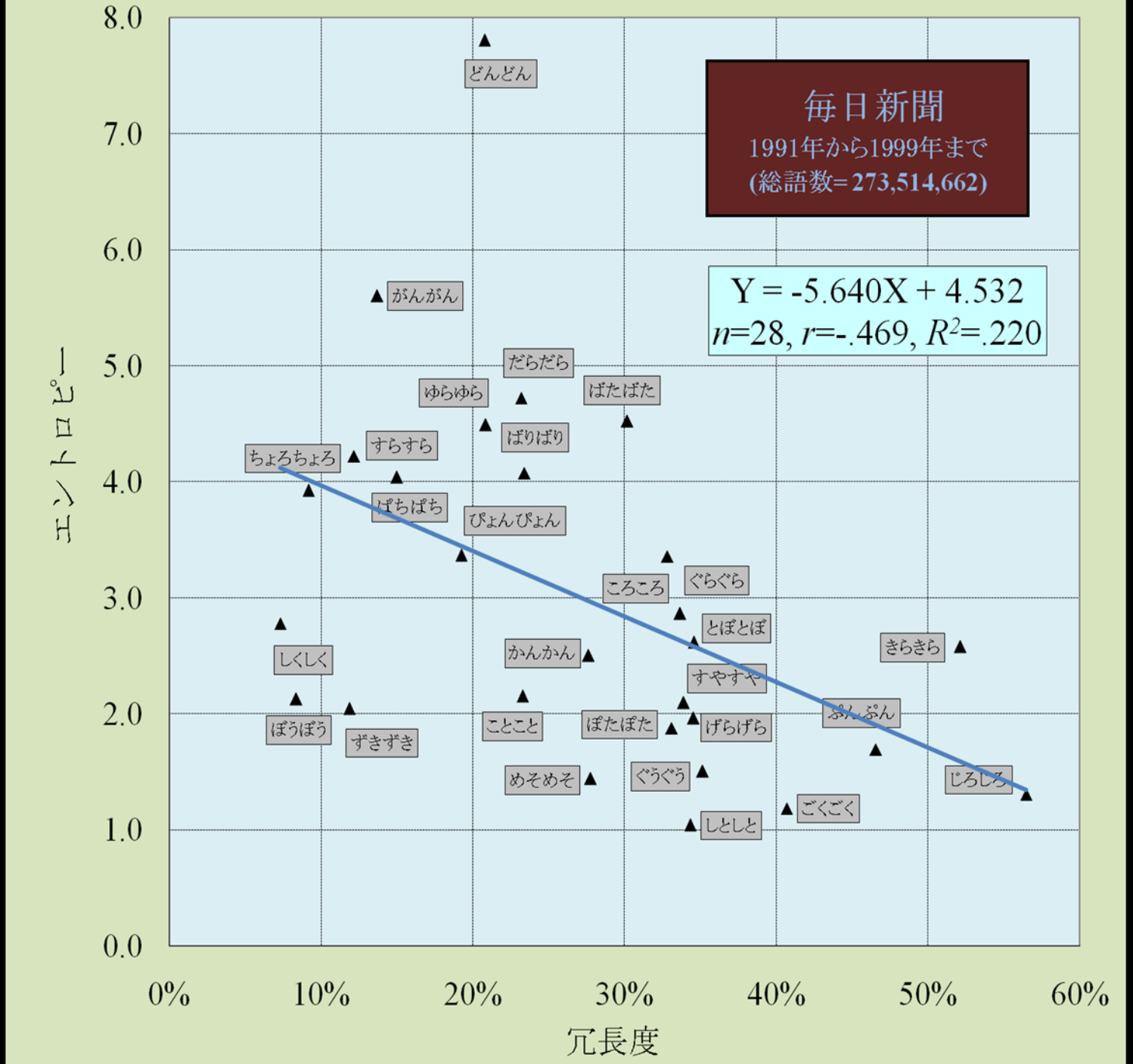
本研究の分析

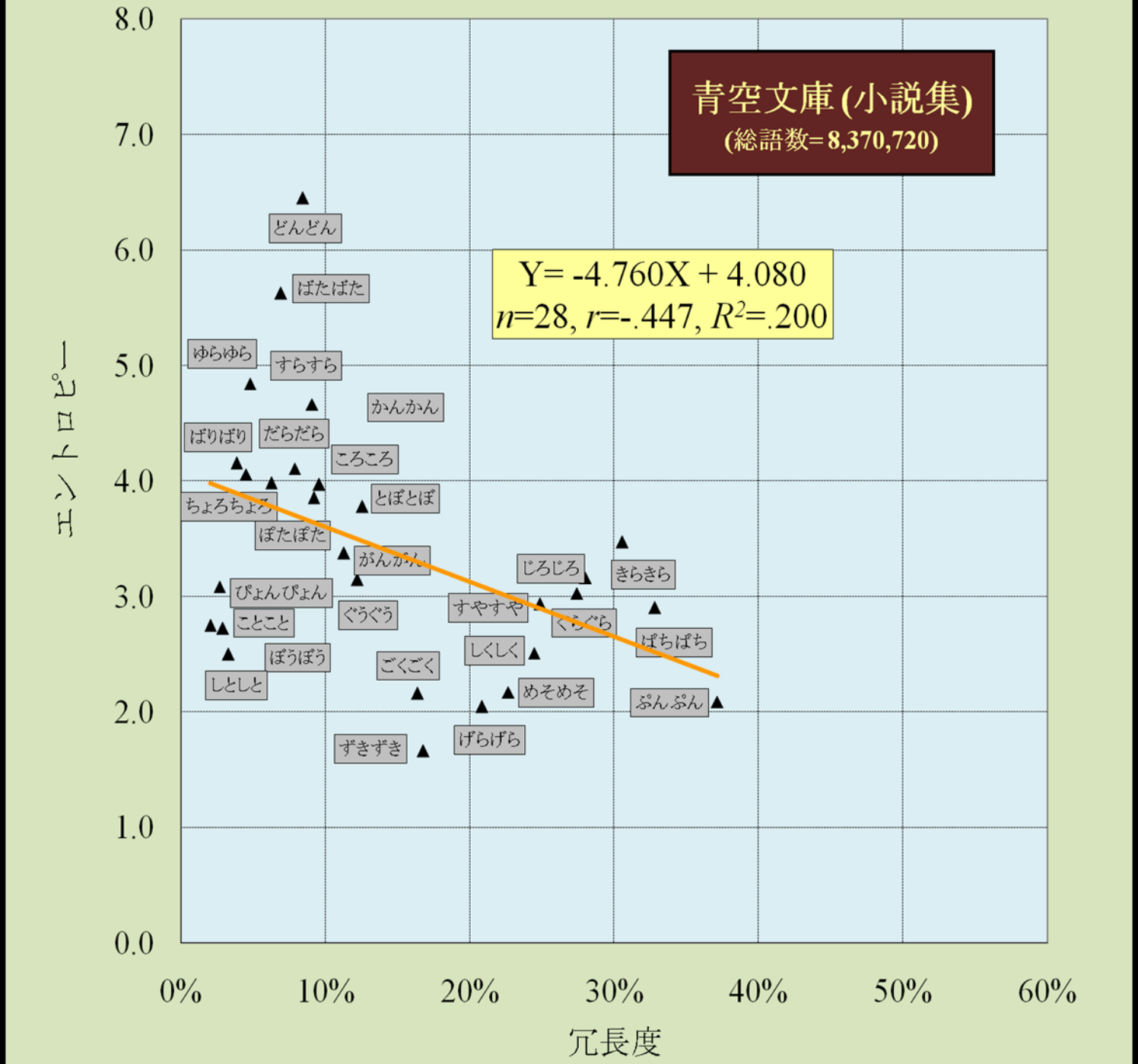
28種類のオノマトペと動詞の共起頻度のパターンについて，**ヒト**，**新聞**，**小説**の3種類のコーパスそれぞれで，エントロピーと冗長度の2つの指標を算出

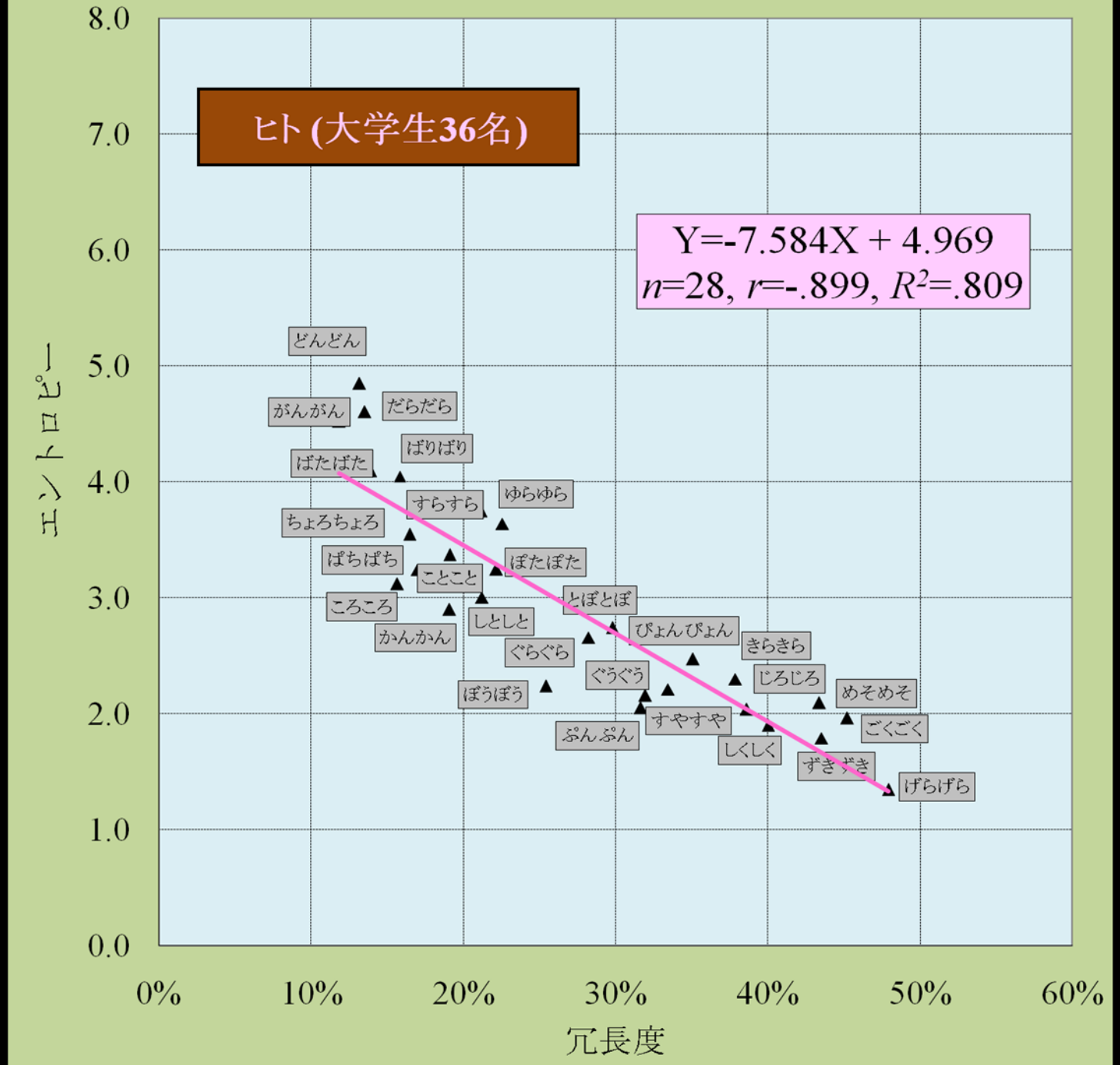
ヒト　**新聞**　**小説**

28オノマトペ

オノマトペ	ヒトの動詞の産出				新聞のコーパスにおける動詞の共起				小説のコーパスにおける動詞の共起			
	動詞種類	産出頻度	エントロピー	冗長度	動詞種類	共起頻度	エントロピー	冗長度	動詞種類	共起頻度	エントロピー	冗長度
どんどん	48	118	4.851	13.14%	926	4424	7.808	20.78%	132	241	6.451	8.43%
だらだら	40	112	4.604	13.49%	71	220	4.723	23.20%	19	28	3.982	6.27%
がんがん	35	83	4.523	11.81%	90	234	5.603	13.69%	14	26	3.377	11.30%
ばたばた	27	77	4.095	13.88%	89	317	4.522	30.17%	66	108	5.627	6.91%
ばりばり	28	80	4.045	15.86%	40	152	4.076	23.41%	20	28	4.155	3.86%
すらすら	27	110	3.750	21.13%	28	62	4.222	12.17%	35	55	4.663	9.08%
ゆらゆら	26	99	3.640	22.56%	51	134	4.490	20.84%	34	49	4.844	4.79%
ちょろちょろ	19	81	3.549	16.45%	20	34	3.925	9.17%	19	26	4.056	4.52%
ことこと	18	52	3.374	19.10%	7	22	2.154	23.29%	7	8	2.750	2.04%
ぱたぱた	18	78	3.246	22.15%	7	19	1.878	33.09%	19	36	3.857	9.20%
ぱちぱち	15	65	3.244	16.98%	27	60	4.042	15.00%	20	79	2.903	32.84%
ころころ	13	63	3.122	15.63%	32	179	3.358	32.83%	21	35	3.971	9.59%
しとしと	14	51	3.005	21.07%	3	12	1.041	34.33%	6	8	2.500	3.29%
かんかん	12	49	2.902	19.06%	11	40	2.504	27.61%	22	43	4.108	7.87%
とぼとぼ	15	62	2.744	29.77%	16	70	2.617	34.56%	20	44	3.778	12.58%
ぐらぐら	13	61	2.657	28.20%	20	87	2.867	33.66%	18	46	3.027	27.42%
ぴょんぴょん	14	91	2.473	35.06%	8	50	3.367	19.25%	9	13	3.085	2.68%
きらきら	13	88	2.300	37.84%	42	442	2.582	52.12%	32	99	3.472	30.56%
ぼうぼう	8	43	2.238	25.41%	5	7	2.128	8.35%	7	9	2.725	2.92%
すやすや	10	59	2.211	33.43%	9	39	2.095	33.89%	15	48	2.935	24.88%
ぐうぐう	9	43	2.158	31.93%	5	13	1.506	35.15%	12	29	3.147	12.21%
めそめそ	13	82	2.096	43.36%	4	23	1.445	27.76%	7	33	2.171	22.66%
ぶんぶん	8	49	2.051	31.65%	9	41	1.694	46.57%	10	27	2.088	37.14%
じろじろ	10	61	2.040	38.59%	8	56	1.306	56.48%	21	96	3.162	28.01%
ごくごく	12	52	1.964	45.21%	4	16	1.186	40.69%	6	10	2.161	16.40%
しくしく	9	42	1.901	40.02%	8	16	2.781	7.31%	10	55	2.510	24.45%
ずきずき	9	68	1.791	43.51%	5	10	2.046	11.86%	4	7	1.664	16.78%
げらげら	6	50	1.346	47.91%	8	31	1.964	34.55%	6	16	2.046	20.84%
平均	17.46	70.32	2.93	26.94%	55.46	243.21	3.00	27.21%	21.82	46.50	3.40	14.27%
標準偏差	10.23	21.44	0.93	10.98%	169.32	811.14	1.52	12.65%	24.60	46.29	1.09	10.26%


毎日新聞
1991年から1999年まで
(総語数=273,514,662)
Y = -5.640X + 4.532
n=28, r=-.469, R²=.220
エントロピー
冗長度
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0%
10%
20%
30%
40%
50%
60%
どんどん
がんがん
だらだら
ゆらゆら
ばたばた
ばりばり
ちょろちょろ
すらすら
ぴょんぴょん
ころころ
ぐらぐら
とぼとぼ
しくしく
かんかん
きらきら
すやすや
ぼうぼう
ずきずき
ことこと
ぽたぽた
げらげら
ぷんぷん
めそめそ
ぐうぐう
じろじろ
ごくごく
しとしと

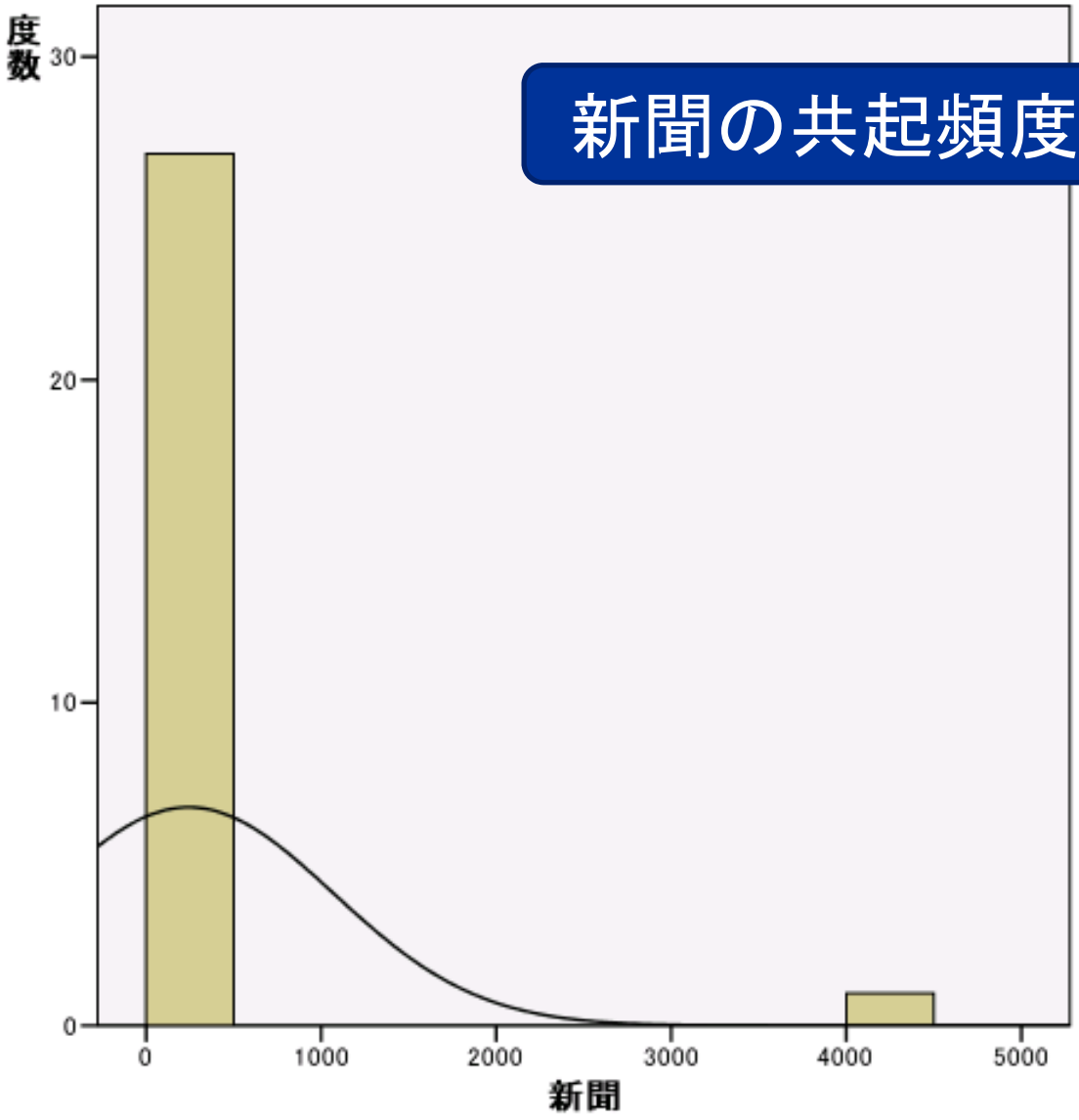

青空文庫 (小説集)
(総語数= 8,370,720)
Y= -4.760X + 4.080
n=28, r=-.447, R2=.200
エントロピー
冗長度
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0%
10%
20%
30%
40%
50%
60%
どんどん
ばたばた
ゆらゆら
すらすら
かんかん
ばりばり
だらだら
ころころ
とぼとぼ
ちょろちょろ
ぽたぽた
がんがん
じろじろ
きらきら
ぴょんぴょん
すやすや
ぐうぐう
ことこと
ぐらぐら
ぱちぱち
しくしく
ぼうぼう
ごくごく
しとしと
めそめそ
ぷんぷん
げらげら
ずきずき


ヒト (大学生36名)
Y=-7.584X + 4.969
n=28, r=-.899, R²=.809
エントロピー
冗長度
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0%
10%
20%
30%
40%
50%
60%
どんどん
がんがん
だらだら
ばりばり
ばたばた
すらすら
ゆらゆら
ちょろちょろ
ぱちぱち
ぽたぽた
ことこと
ころころ
とぼとぼ
しとしと
かんかん
ぴょんぴょん
きらきら
ぐらぐら
ぐうぐう
じろじろ
ぼうぼう
めそめそ
すやすや
ごくごく
ぷんぷん
しくしく
ずきずき
げらげら

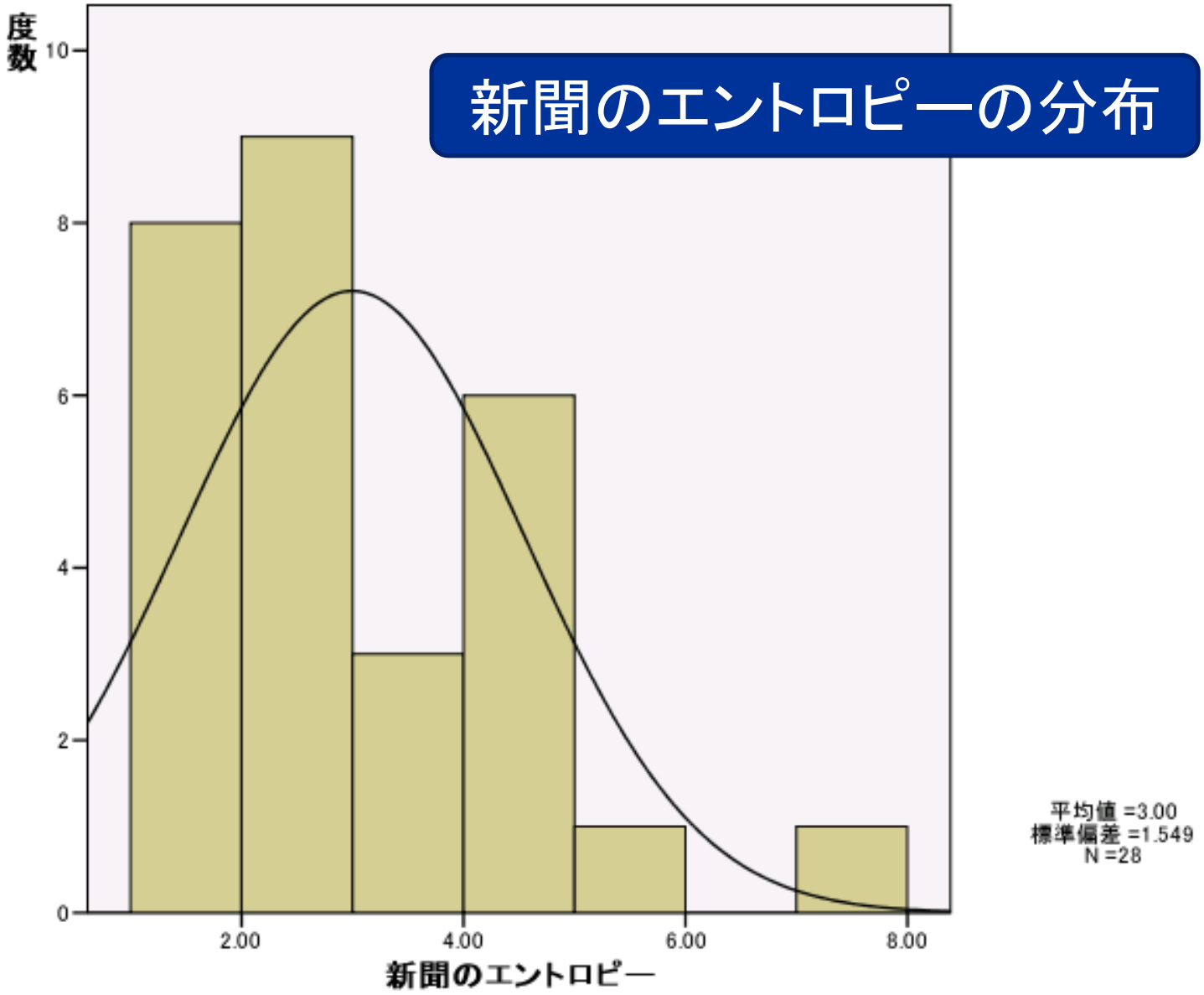
エントロピーと冗長度

- エントロピーと冗長度が同じものを計っているのではないかという疑問があろう. (同じ指標)
- オノマトペと動詞の共起頻度の場合は，毎日新聞で－0.469，青空文庫で－0.447であり，絶対値で0.500を超えていないことから，測定しているものは違っていると考えるのが妥当であろう. (ただし，ヒトの場合は相関が高かった.)
- また，複動動詞の分析(Tamaoka, Lim & Sakai, 2004)で，エントロピーと冗長度で異なる結果が出ていることからも，別の指標であると考えられる.

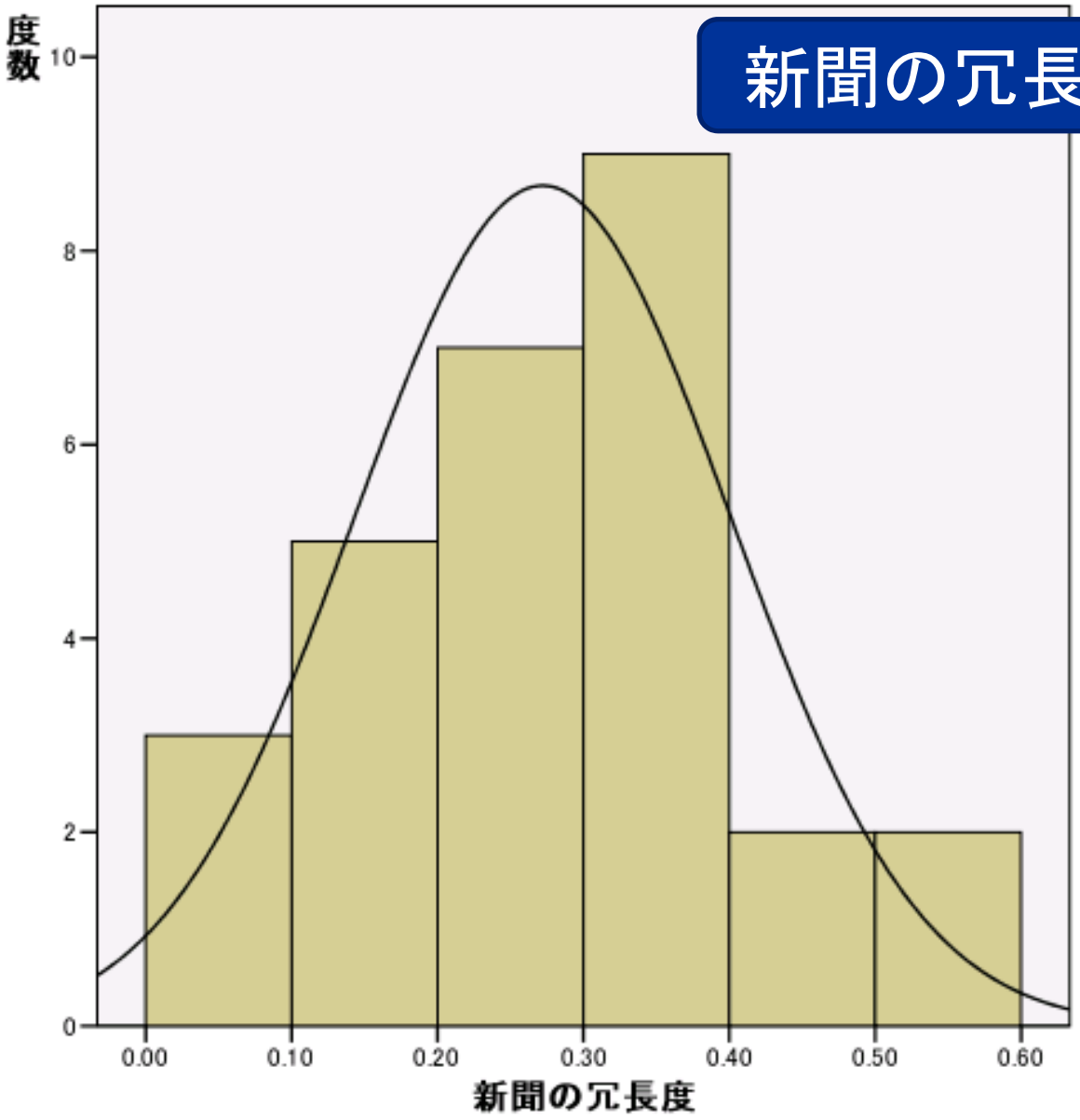
新聞の共起頻度の分布


度数
30
20
10
0
0
1000
2000
3000
4000
5000
新聞
平均値 =243.21
標準偏差 =826.022
N =28

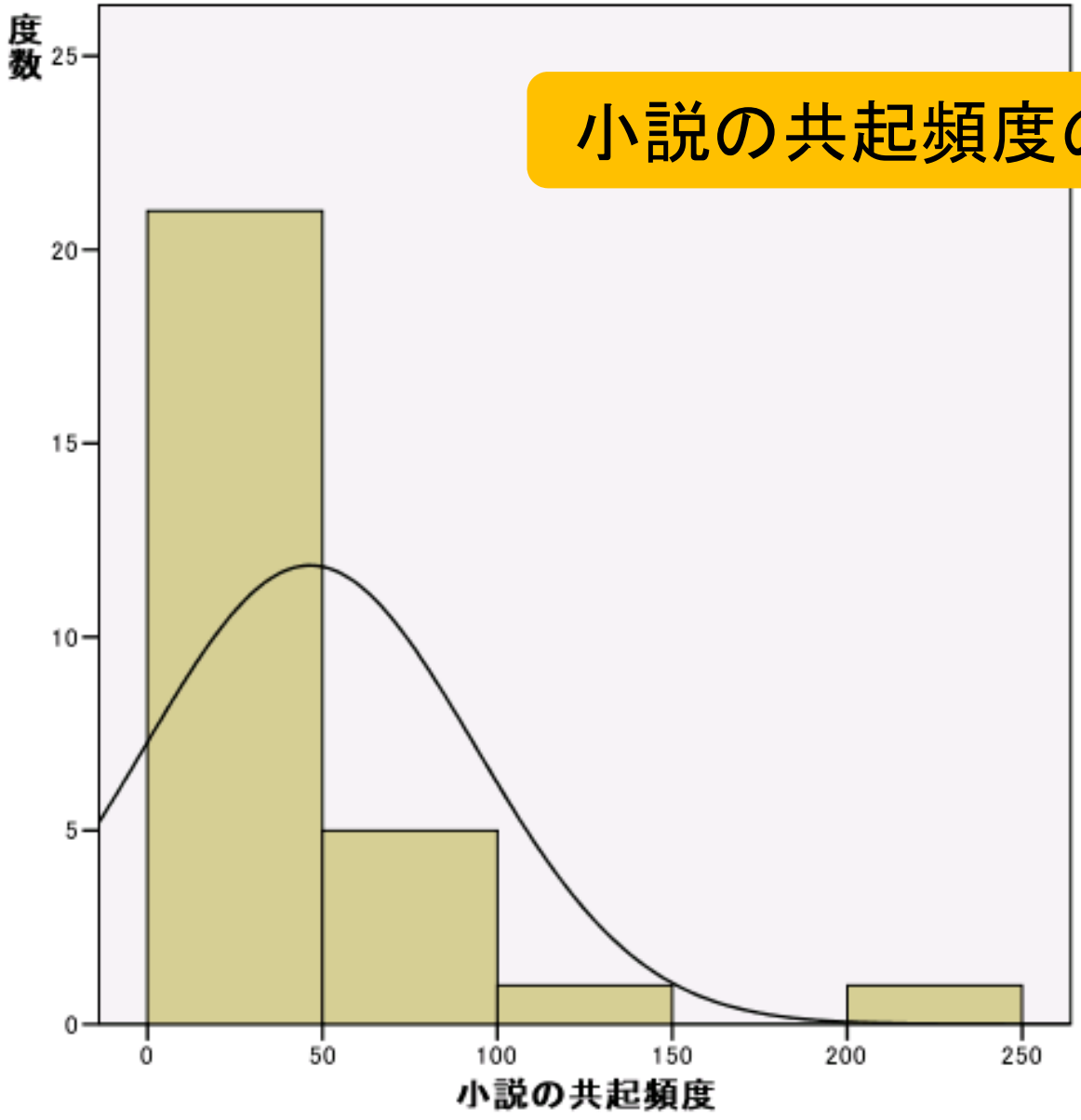
新聞のエントロピーの分布



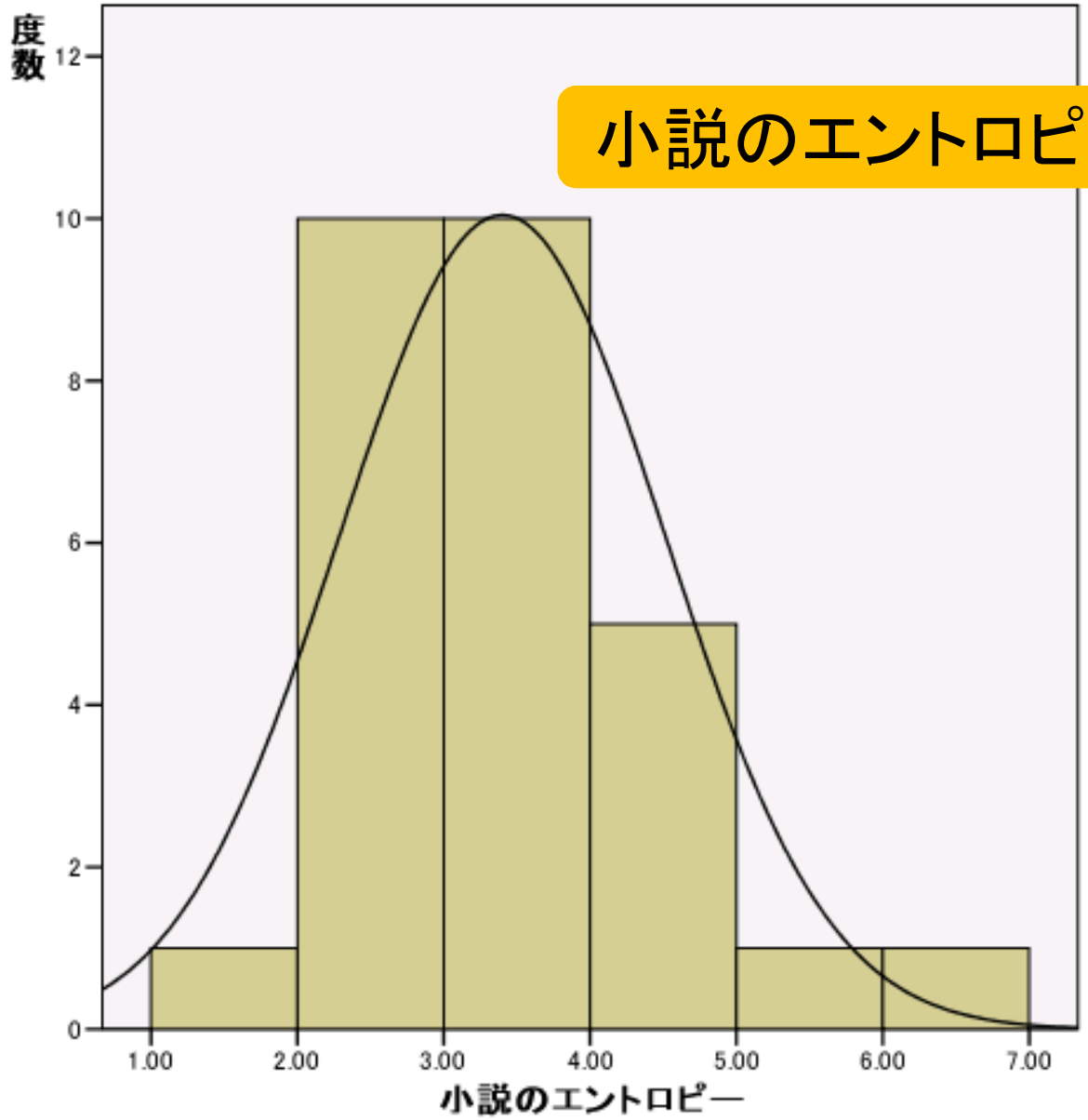
新聞の冗長度の分布


度数
10
8
6
4
2
0
0.00
0.10
0.20
0.30
0.40
0.50
0.60
新聞の冗長度
平均値 =0.27
標準偏差 =0.129
N =28

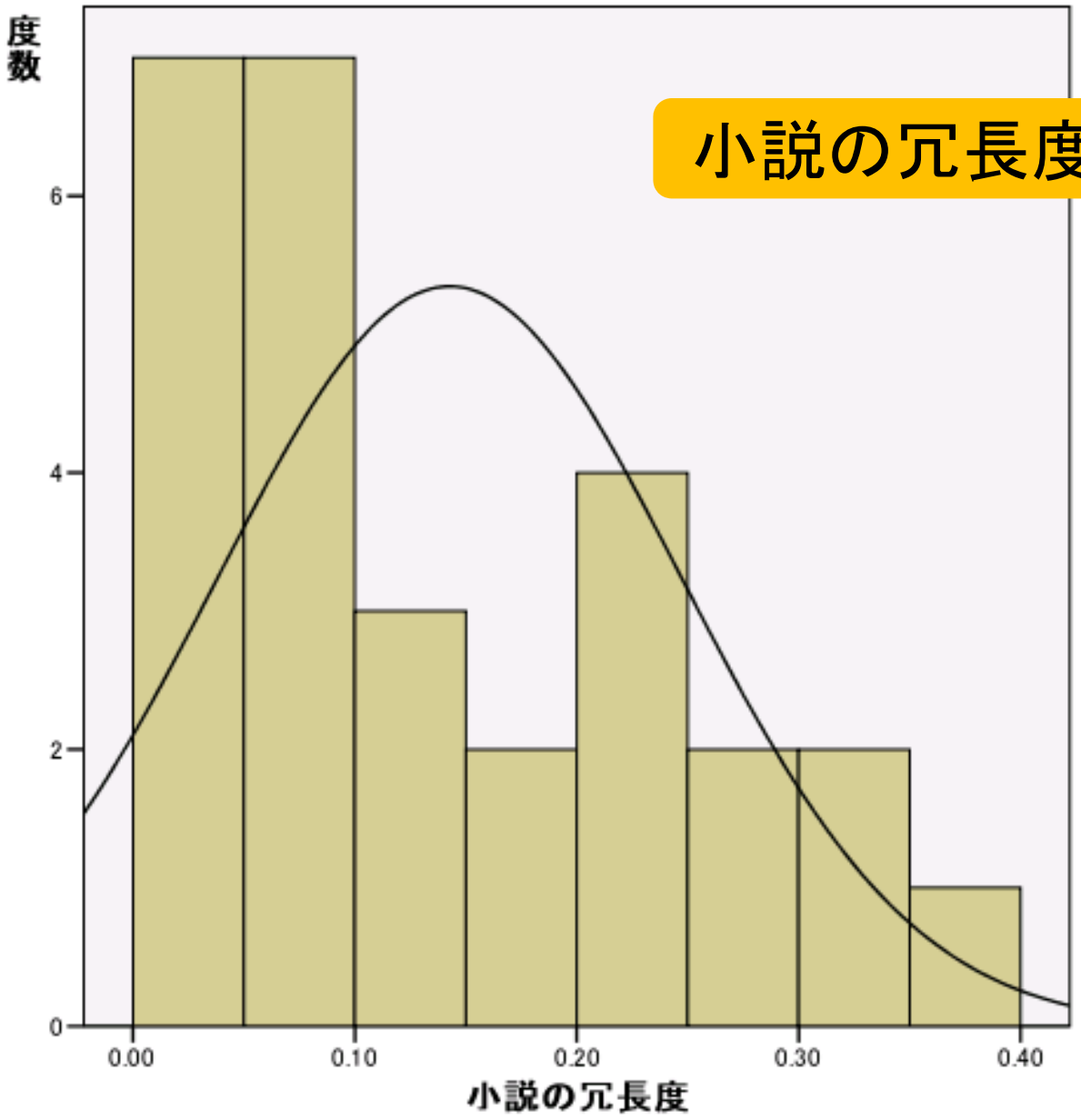
小説の共起頻度の分布


度数
25
20
15
10
5
0
0
50
100
150
200
250
小説の共起頻度
平均値 =46.5
標準偏差 =47.14
N =28

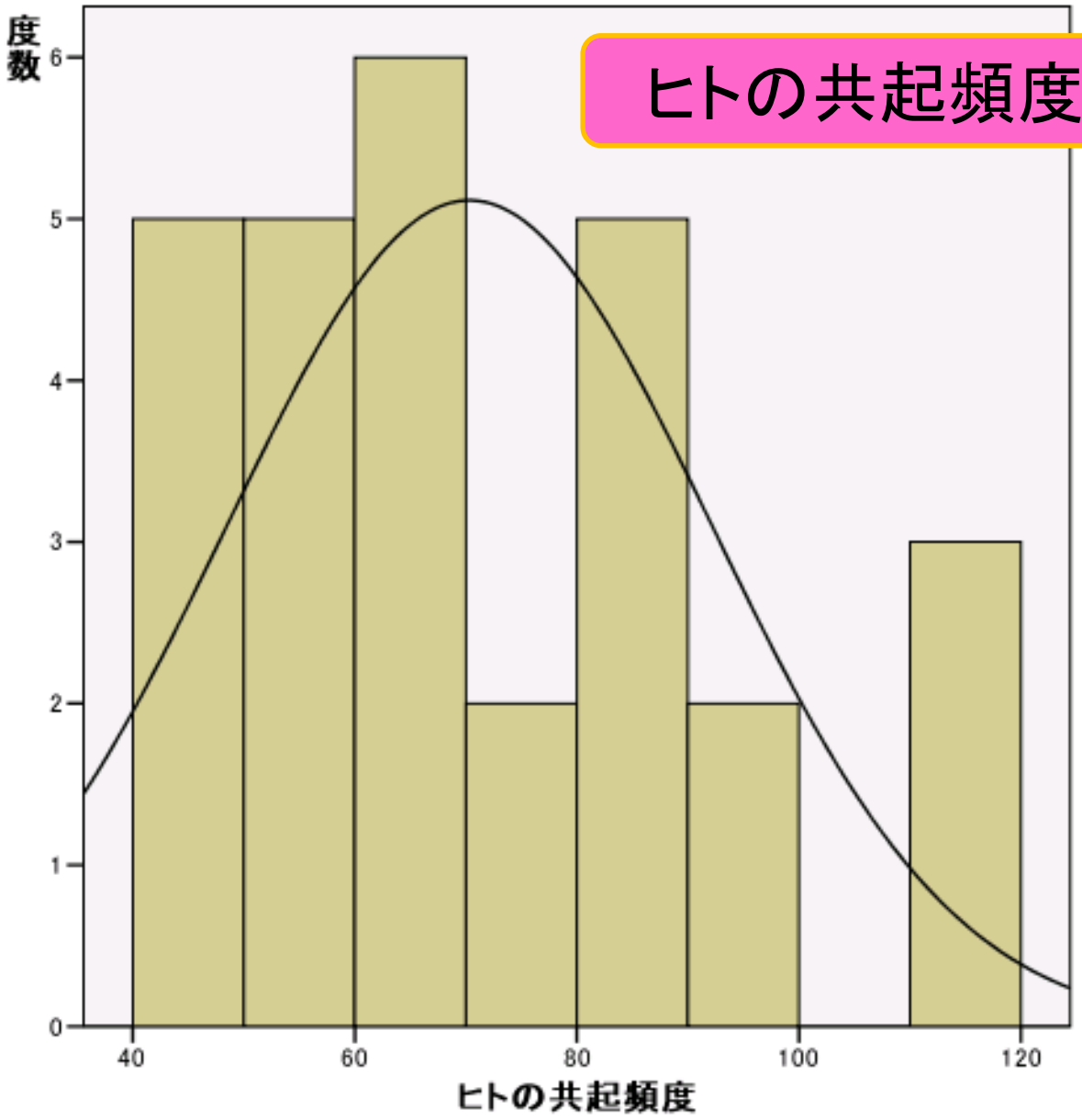
小説のエントロピーの分布



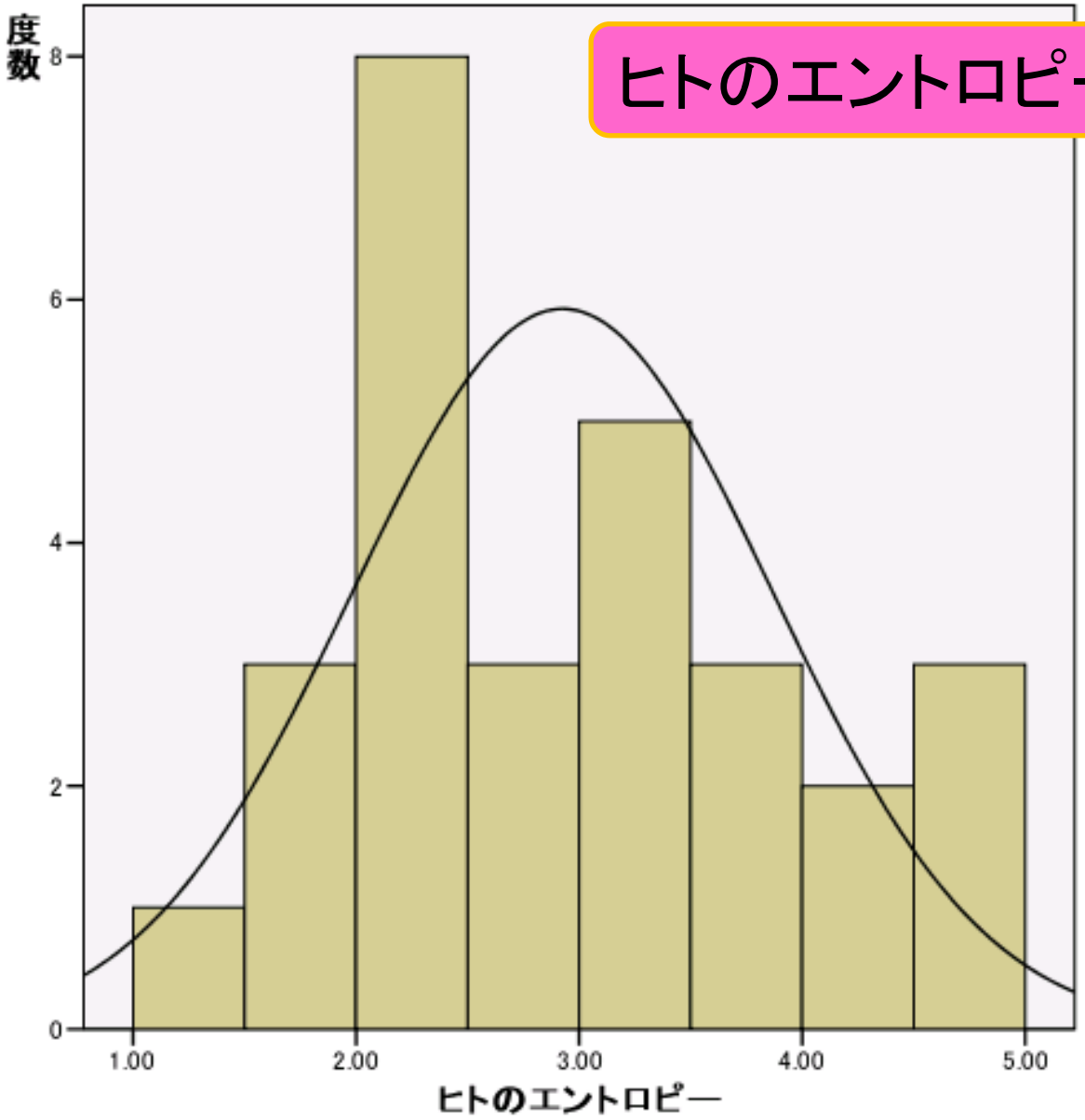
小説の冗長度の分布



ヒトの共起頻度の分布

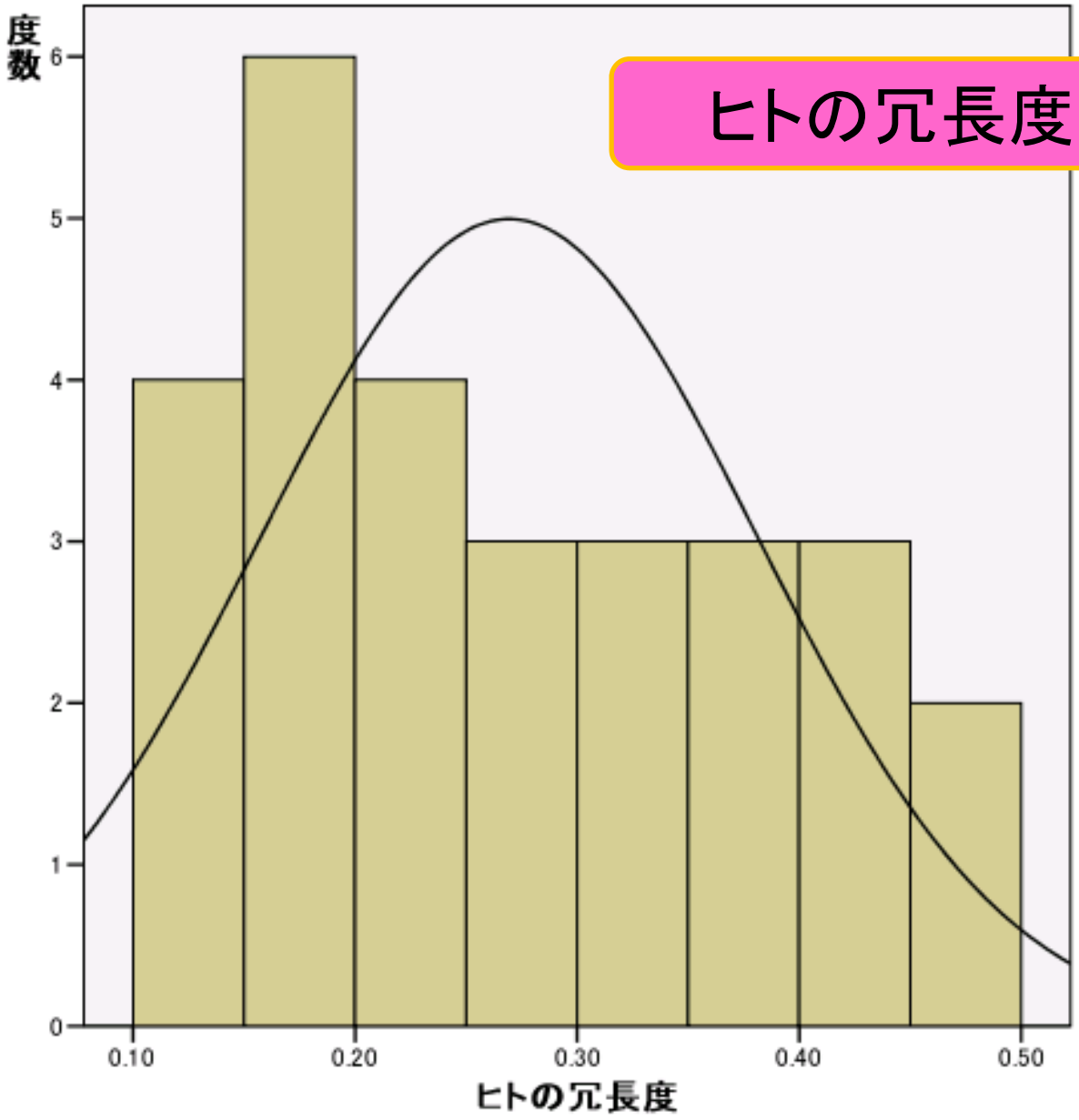


ヒトのエントロピーの分布



平均値 =2.93
標準偏差 =0.943
N =28

ヒトの冗長度の分布



分布と統計手法ー新聞の例

- オノマトペの共起頻度は，正規分布とは程遠いので，共起頻度をそのままパラメトリック統計で解析することはできない。
- エントロピーと冗長度とは，パラメトリック統計で解析できるような分布を示している。

エントロピーの指標によるコーパス頻度とヒトの産出の類似性の検討

表2 エントロピーの相関及び分散分析

	データの出典	1	2	3
1	ヒトの産出	—		
2	新聞のコーパス	0.829 ***	—	
3	小説のコーパス	0.783 ***	0.773 ***	—
	平均	2.926	2.998	3.401
	標準偏差	0.943	1.549	1.112
	分散分析の結果	$F(2,54)=4.735, p<.05$		
	単純対比の結果	ヒト＝新聞＜小説		

注: n(オノマトペの数)=28. *** $p<.001$.

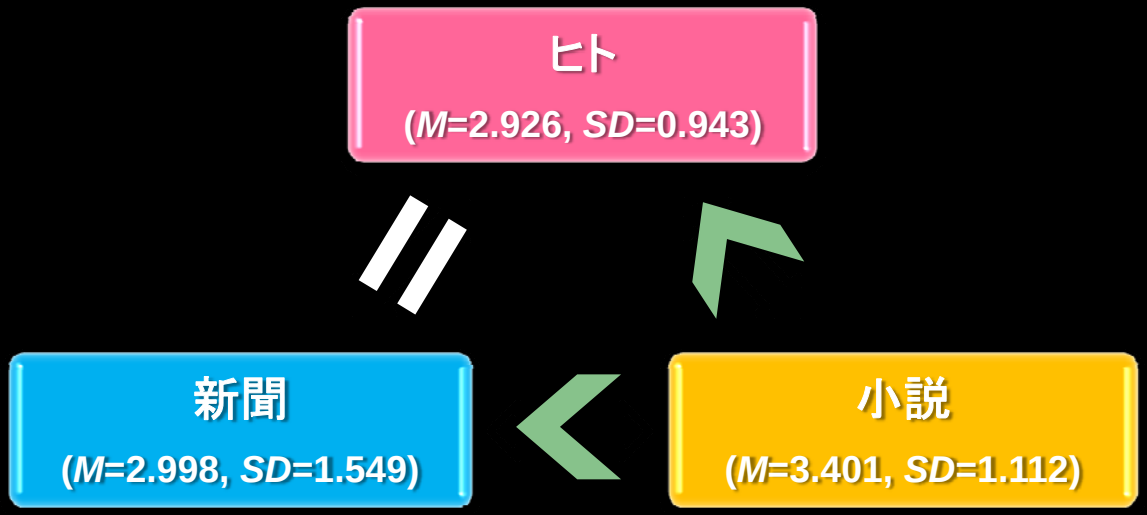
3種類のコーパス間の相関(ピアソンの相関係数)はすべて高い。

3種類のコーパスのエントロピーの分散分析の結果

- コーパスの種類について有意な主効果

  [$F(2,54)=4.735$, $p<.05$]

- シェフェの多重比較



ヒト＝新聞＜小説

冗長度の指標によるコーパス頻度とヒトの産出の類似性の検討

表3　冗長度の相関及び分散分析

	データの出典	1	2	3
1	ヒトの産出	—		
2	新聞のコーパス	0.346	—	
3	小説のコーパス	0.523 **	0.455 *	—
	平均	26.935%	27.206%	14.268%
	標準偏差	11.180%	12.879%	10.446%
	分散分析の結果	$F(2,54)=20.146, p<.001$		
	単純対比の結果	ヒト＝新聞＜小説		

注: n(オノマトペの数)=28.　* $p<.05$.　** $p<.01$.

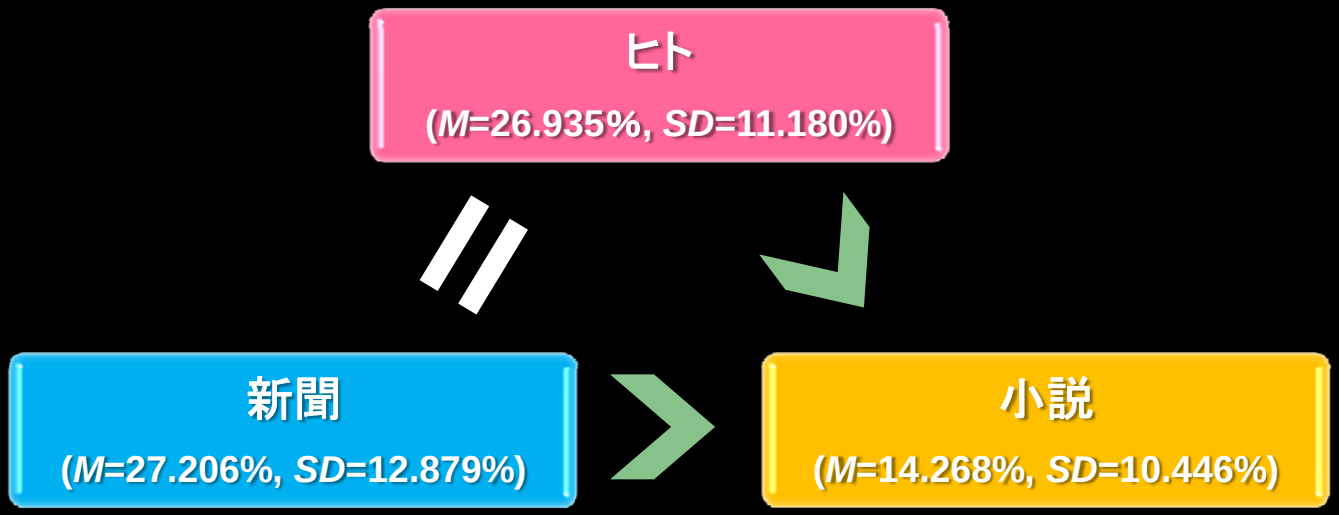
ヒトと小説、新聞と小説の相関は高い。

3種類のコーパスの冗長度の分散分析の結果

- コーパスの種類について有意な主効果

  [$F(2,54)=20.146$, $p<.001$]

- シェフェの多重比較



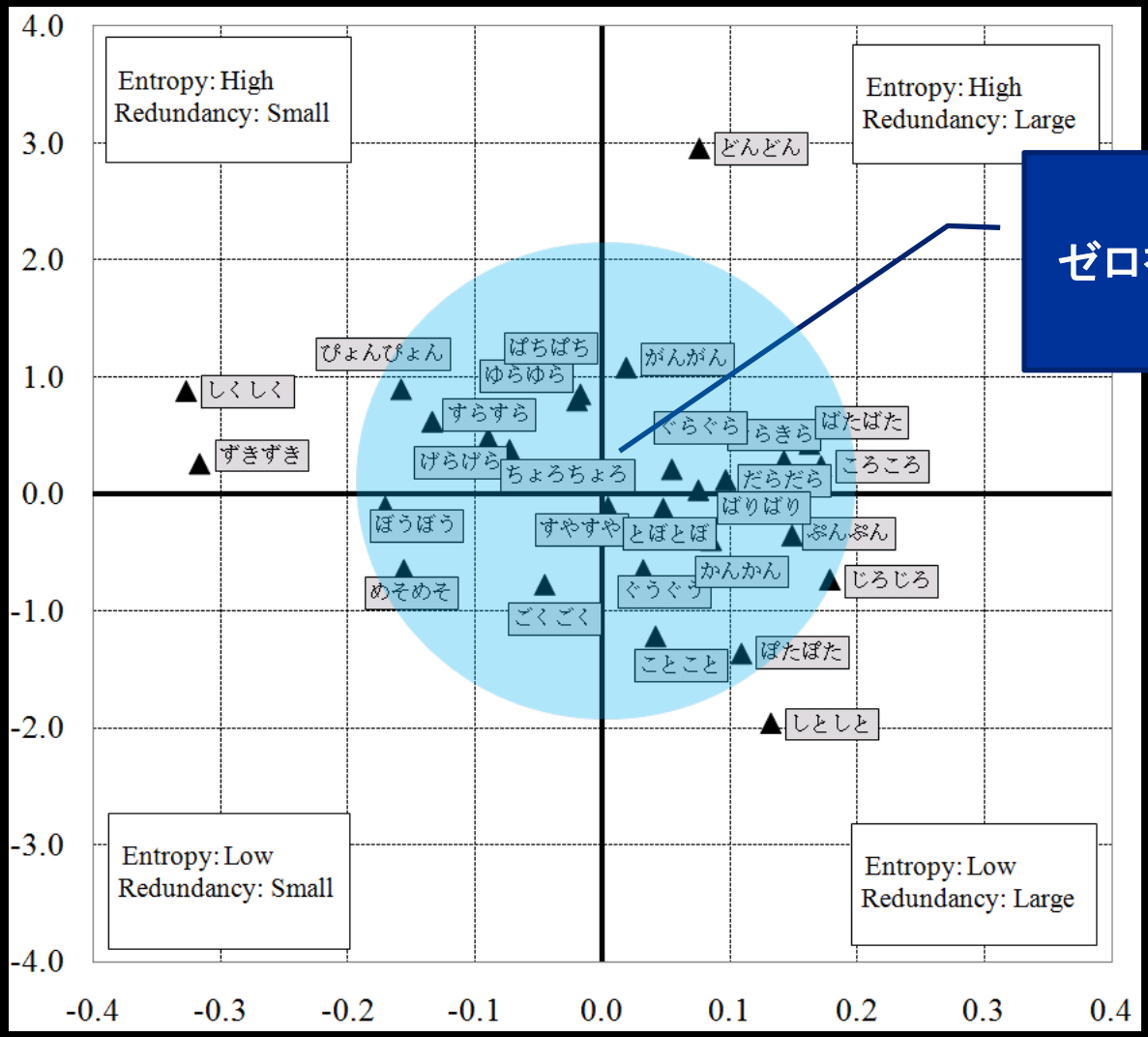
ヒト＝新聞＞小説

結果のまとめ

28種類のオノマトペと共起する動詞の頻度パターンについて，新聞と小説，及びヒト（大学生）を対象とした産出テストのデータという3種類のコーパスを用いて比較した。

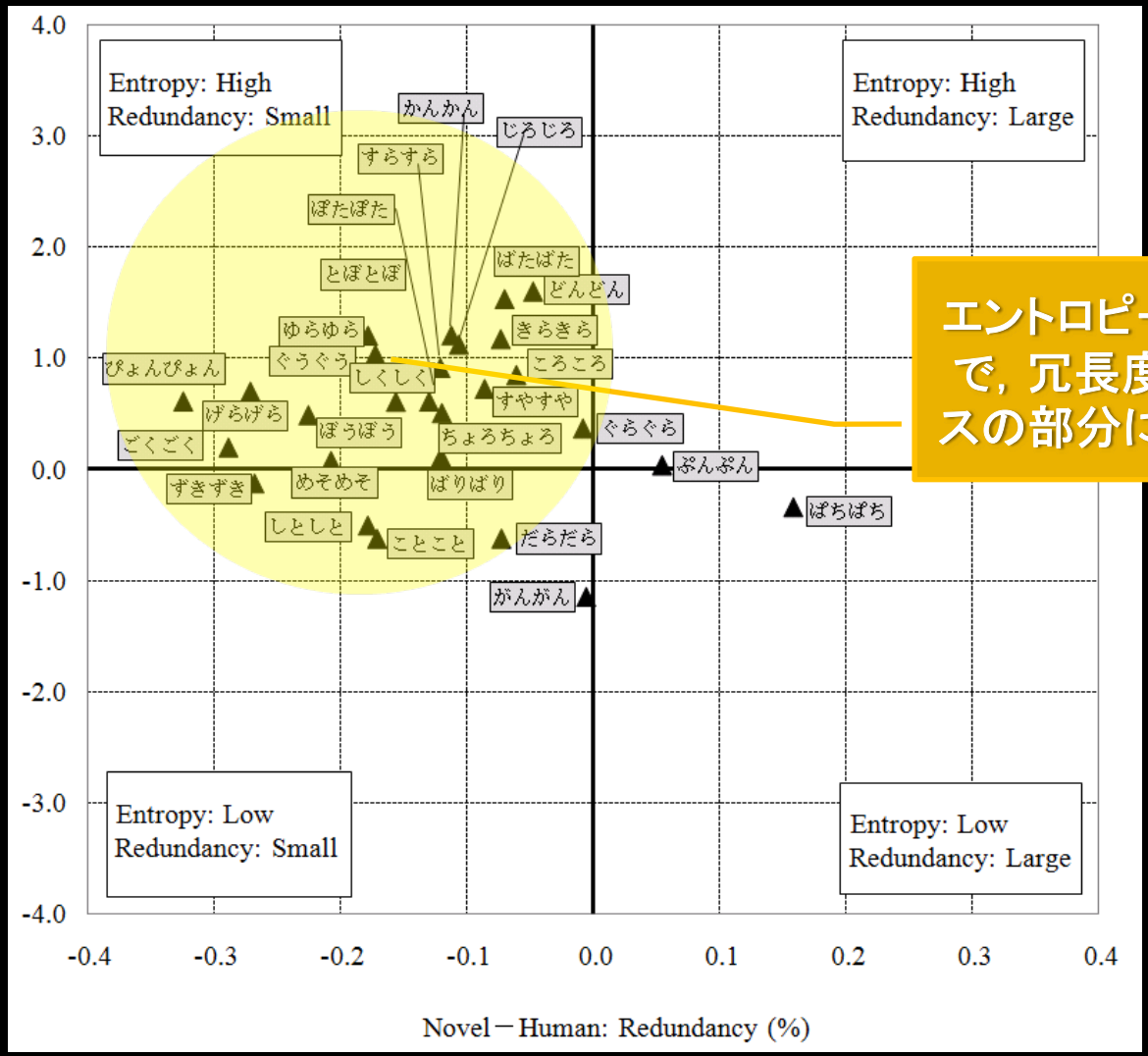
- 新聞のコーパスから得られた動詞の頻度パターンはヒトの産出と類似している。
- 小説のコーパスとヒトの産出には大きな違いがみられた。

エントロピーと冗長度に関する
新聞とヒトの差



ゼロを中心に分布

エントロピーと冗長度に関する
小説とヒトの差



エントロピーの差はプラスで，冗長度の差はマイナスの部分に大多数が分布

小説－ヒト：エントロピー

小説－ヒト：冗長度（%）

結果の考察

- 新聞は複数の新聞記者が一般大衆に情報を伝達するために，簡潔で分かりやすい表現を目指している

←相対的に，エントロピーが低く，冗長度が高い

- 小説が特定の作家の個性によって多様な表現が駆使されている

←相対的に，エントロピーが高く，冗長度が低い

研究の結論

オノマトペと動詞の共起表現についてのみの限られた知見

小説のコーパスに基づいてヒトの言語産出を一般化して論ずるのは難しい。むしろ新聞のコーパスの方が，ヒトの言語産出を適切に反映していると言えるようである。

エントロピーと冗長度を使った研究－1

Tamaoka, K., Lim, H., & Sakai, H. (2004). Entropy and redundancy of Japanese lexical and syntactic compound verbs. *Journal of Quantitative Linguistics*, *11(3)*, 233-250.

エントロピーと冗長度で，語彙的複合動詞と統語的複合動詞の特徴を比較検討した研究である．コーパスは，毎日新聞と青空文庫を使用している．

エントロピーと冗長度を使った研究－2

玉岡賀津雄・宮岡弥生・林炫情 (2003).エントロピーと冗長度で表現の多様性と規則性を表す試みー 韓国語系日本語学習者の敬語表現を例に. *日本語科学*, *14*, 98-112.

韓国語を母語とする日本語学習者による書く条件と話す条件の尊敬と謙譲表現の多様性をエントロピーと冗長度で測定して比較した研究である. コーパスは, 先生とのインタヴュー場面を設定して独自に作成している.

エントロピーと冗長度を使った研究－3

Miyaoka, Y., & Tamaoka, K. (2005). An Investigation of the *Right-hand Head Rule* Applied to Japanese Affixes. *Glottometrics, 10, 45-54*

漢字一字で書かれる接頭辞と接尾辞の違いを，1985年から1998年の14年間の朝日新聞の語彙コーパスを使って検討した研究である．接頭辞の方が接尾辞よりもエントロピーが高く，接頭辞は接尾辞よりもより不規則に名詞に付加されていることを示し，右側主要部の規則（right-hand head rule)を支持する結果を得ている．

エントロピーと冗長度を使った研究－4

玉岡賀津雄・木山幸子・宮岡弥生(2008). ヒトの言語産出とコーパスの頻度はどのくらい類似しているか, *日本言語学会第136回大会予稿集*(学習院大学), 122-127.

- エントロピーと冗長度で，オノマトペと動詞の共起パターンを，新聞，小説，ヒトの３種類で比較した研究である．ヒトと新聞とが類似した共起パターンを示し，小説は両者と異なっていた．

引用文献ーエントロピー関係

有本卓 (1982). *確率・情報・エントロピー*. 東京: 森北出版.

堀淳一 (1979). *エントロピーとは何か*. 東京: 講談社ブルーバックス

海保博之 (1989). 第1講: 情報をはかるーエントロピー・情報伝達量・冗長度. 海保博之 (編), *心理・教育データの解析法10講ー応用編 (pp.14-26)*. 東京: 福村出版

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379-423 (Part I) and 623-656 (Part II).

他にもあるよ！ー他の分析法1
二項ロジステック回帰

Tamaoka, K., Matsuoka, C., Sakai, H., & Makioka, S. (2005). Predicting attachment of the light verb – *suru* to Japanese two-kanji compound words using four aspects. *Glottometrics, 10*, 73-81.

漢語名詞(「故障」,「接続」,「連絡」など) を「終結」,「持続」,「開始」および「状態」の4種類のアスペクトに分類し，スル軽動詞（サ変動詞）の結合を予測した.「終結」の予測率は，スル軽動詞が付加される802語のうち751語で，93.64%であり，エラーはわずかに8語で，1.05%であった. さらに,「終結」のアスペクトは他の「持続」や「開始」のアスペクトと重なることが多いことも示した.

他にもあるよ！ー他の分析法2
「決定木（decision tree）」分析

玉岡賀津雄 (2006).「決定木」分析によるコーパス研究の可能性: 副詞と共起する接続助詞「から」「ので」「のに」の文中・文末表現を例に. *自然言語処理, 13*(2), 169-179.

「決定木」分析を使って，3種類の接続助詞「から」「ので」「のに」が，7種類の副詞「何しろ」「何せ」「せっかく」「現に」「どうせ」「実際」「本当に」と共起する場合に，文中と文末の表現でどちらが使われるかを，新潮文庫100冊のコーパスから得た共起頻度を使って解析した. ノンパラメトリックの多変量解析の一種である.

他にもあるよ！ー他の分析法3
チェビシェフの不等式

Tamaoka, K. Makioka, S. & Murata, T. (2004). Are the effects of vowel repetition influenced by frequencies?: A corpus study on CVCVCV-structured nouns with and without vowel. *Glottometrics*, *8*, 1-11.

CVCVCV構造の日本語の名詞の母音反復の頻度を天野・近藤(2000)の朝日新聞の語彙頻度データベースを使って検討した．CVが3つ連続する条件で3つの母音がすべて同じになるランダム確率は4%であるが，それを大きく超えて9.15%となった．チェビシェフの不等式を使って確率を計算した結果，これは有意に高い出現頻度であった．

他にもあるよ！ー他の分析法4
対称性(symmetry)の検討

Tamaoka K., & Altmann, G. (2004). Symmetry of Japanese kanji lexical productivity in the left- and right-hand sides. *Glottometrics*, *7*, 65-84.

常用漢字1,945字について，左右の熟語生成が対称性を示すかどうかを検討した．分析の結果，個々の漢字のレベルでは46.38%が左右対称，20.72%が左側に歪んでおり，21.23%が右側に歪んでおり，残りの11.67%は熟語生成数が少ないので計算ができなかった．