文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築: 21世紀の日本語研究の基盤整備」公開ワークショップ

エントロピーと冗長度を指標とした語彙的・統語的複合動詞の比較研究

玉岡 賀津雄 (TAMAOKA, Katsuo)
麗澤大学 (Reitaku University, Japan)

東京工業大学大岡山キャンパス・ディジタル多目的ホール, 2009年3月16日(月) 10:00-1200

1．共起頻度の分析のための指標

Entropy (エントロピー)

あいまいさや乱雑さの増減を示す指標

多様性

Redundancy (冗長度)

エントロピーとエントロピー最大値から得られる
無駄の程度を示す指標

規則性

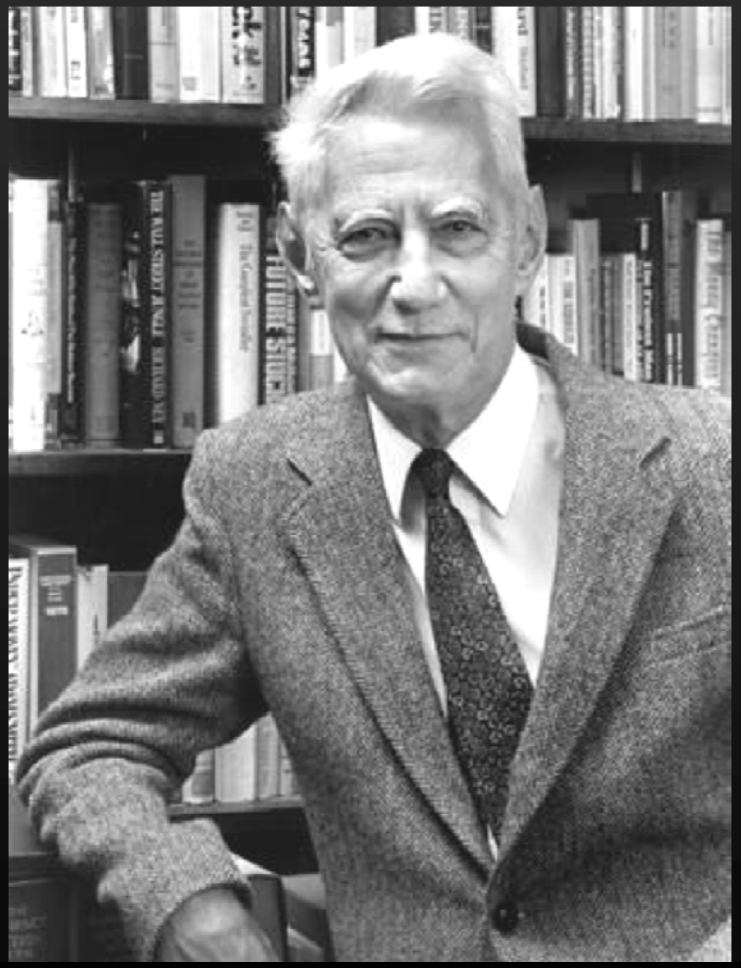
Claude Shannon: エントロピーと冗長度を提案した

デジタル回路の
数学的基礎を確立

1916年
米ミシガン州に
生まれる

ミシガン大学
数学と電気工学
の学士号

MIT
数学と電気工学
の博士号を同時
に取得 (**1941年**)



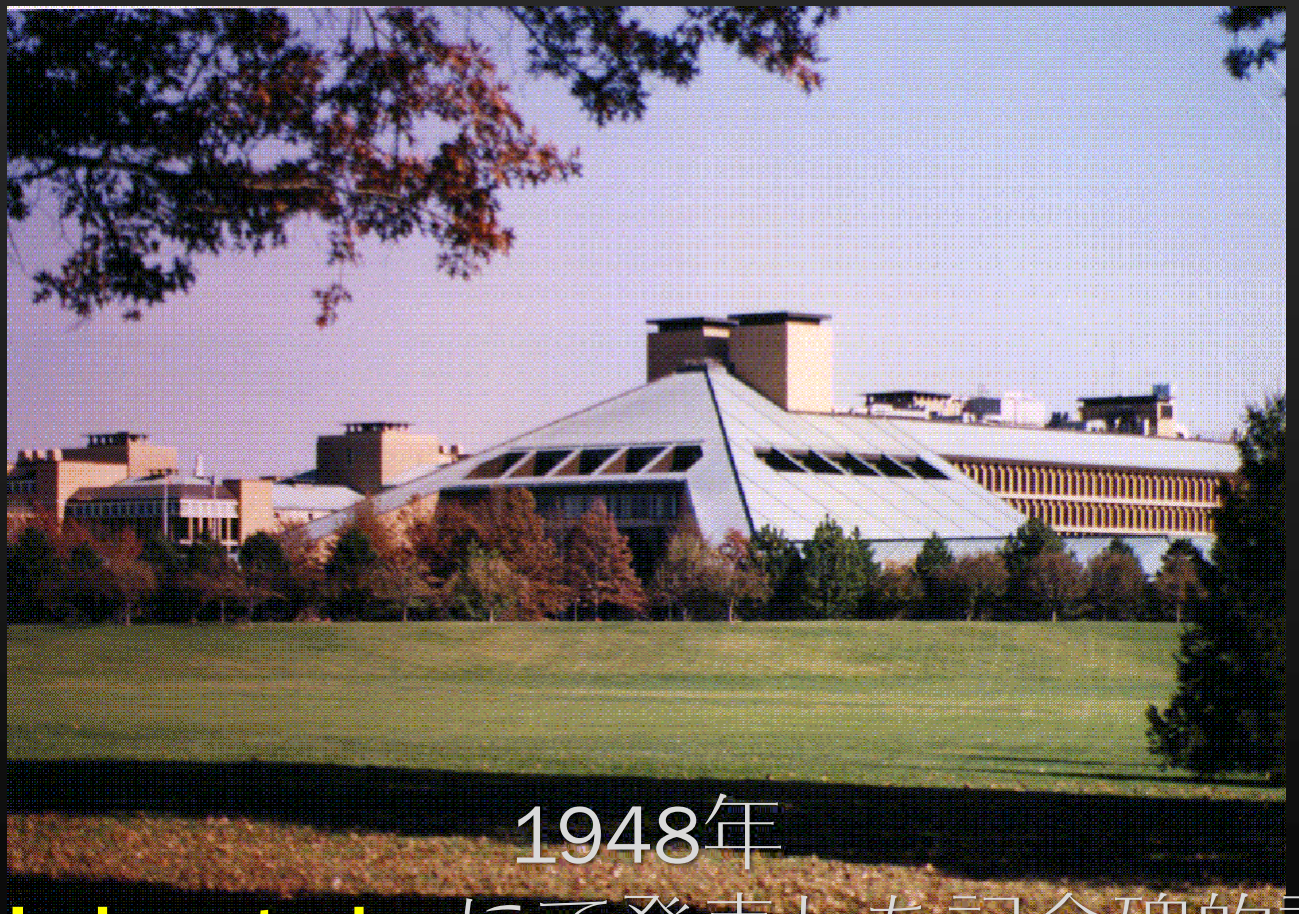
1941年
Bell Laboratories
に就職

MITへ
1956年客員教授
1958年教授に

2001年2月26日
84歳で死去

Claude Shannon

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



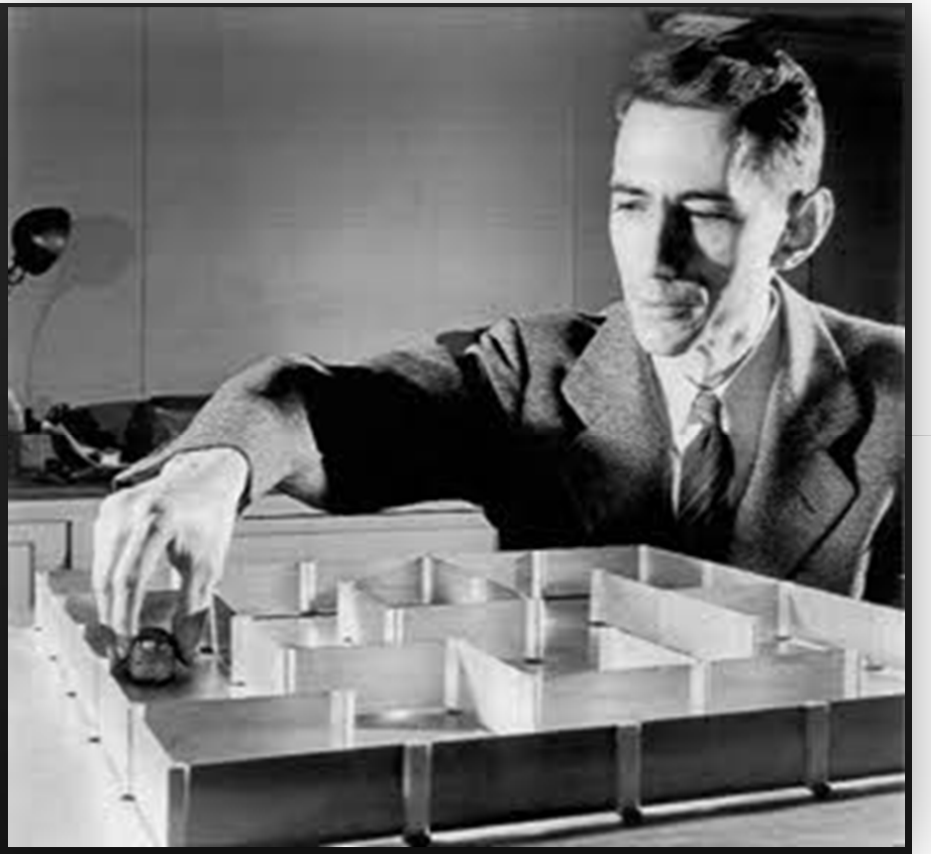
1948年
Bell Laboratoriesにて発表した記念碑的論文

Claude Shannon

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

Claude Shannon

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。



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る。

エントロピーと冗長度を用いたコーパス研究

Tamaoka, K., Lim, H., & Sakai, H. (2004).

Entropy and redundancy of Japanese lexical and syntactic compound verbs.

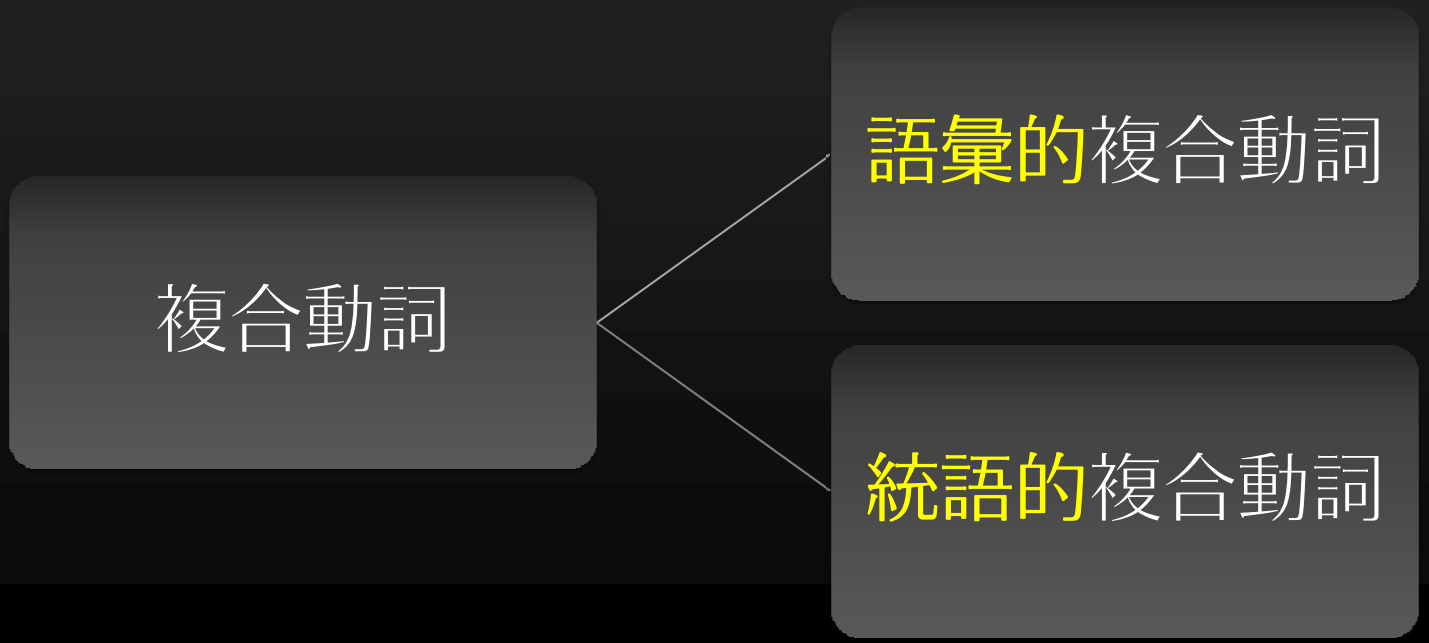
Journal of Quantitative Linguistics, *11*, 233-250.

本発表では，エントロピーと冗長度の
使用の有用性に焦点を当てて説明する。

二種類の複合動詞

- 日本語では，動詞を二つ組み合わせることで複合動詞を作ることができる。

- 影山 (1993, 1999)
  日本語の複合動詞を 2 種類に分けている (影山, 1999, p. 189)。



2．語彙的・統語的複合動詞の統語構造－その 1

語彙的複合動詞　(lexical compound verbs)

- 初めにくる動詞(V_1)と次にくる動詞(V_2)の組み合わせに「語彙的な習慣化が見られる」（影山, 1999, p. 189）

- 語彙的複合動詞には，二つの動詞の間に「語の形態的親密性」を示すサエやモを挿入することができず，一つの語となっていると指摘している。

2．語彙的・統語的複合動詞の統語構造－その 2

統語的複合動詞　(syntactic compound verbs)

- 影山（1999）
  統語的複合動詞はその名の示す通り「統語的な構造」に由来し，補文構造という形で捉えられる (pp. 189-190)

例:「彼は昼食を食べ始めた。」と「子供は手紙を投函し忘れた。」における補文構造は，「彼は [昼食を食べ] 始めた。」と「子供は [手紙を投函し] 忘れた。」という「V1することを(が)V2」の部分を指す。

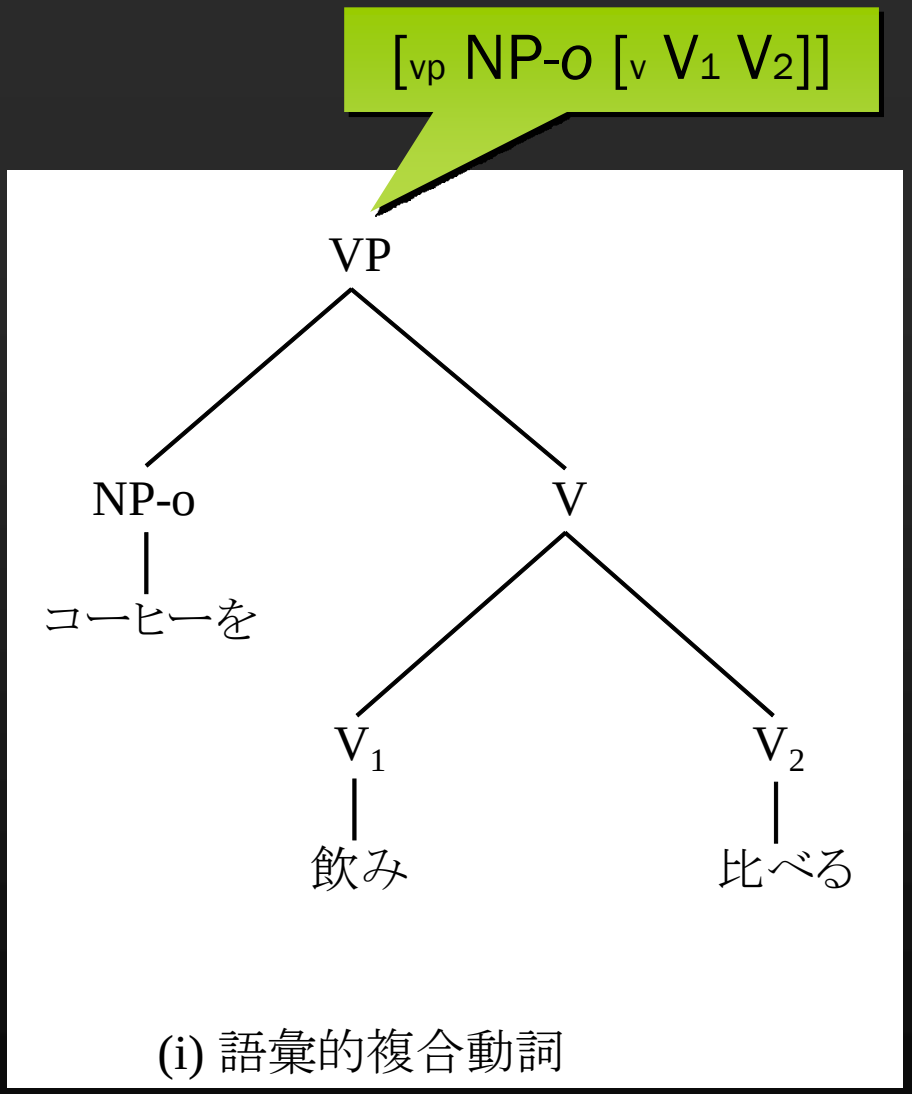
- この構造をもとにして，「食べ」を「始める」に，「投函し」を「忘れる」に結合させると，表面的には，「食べ始める」と「投函し終わる」という複合動詞になると説明している。

語彙的複合動詞の統語構造

語彙的複合動詞

先にくる動詞(V_1)である「飲む」が，次にくる動詞(V_2)の「比べる」と結合して「飲み比べる」という一つの動詞 ($V_1 + V_2 = V$) を作る。

この複合動詞に対して，「コーヒーを」という対格の名詞句(NP, noun phrase)が結びついて動詞句(VP, verb phrase)を作るという構造を持つ。



(i) 語彙的複合動詞

統語的複合動詞の統語構造

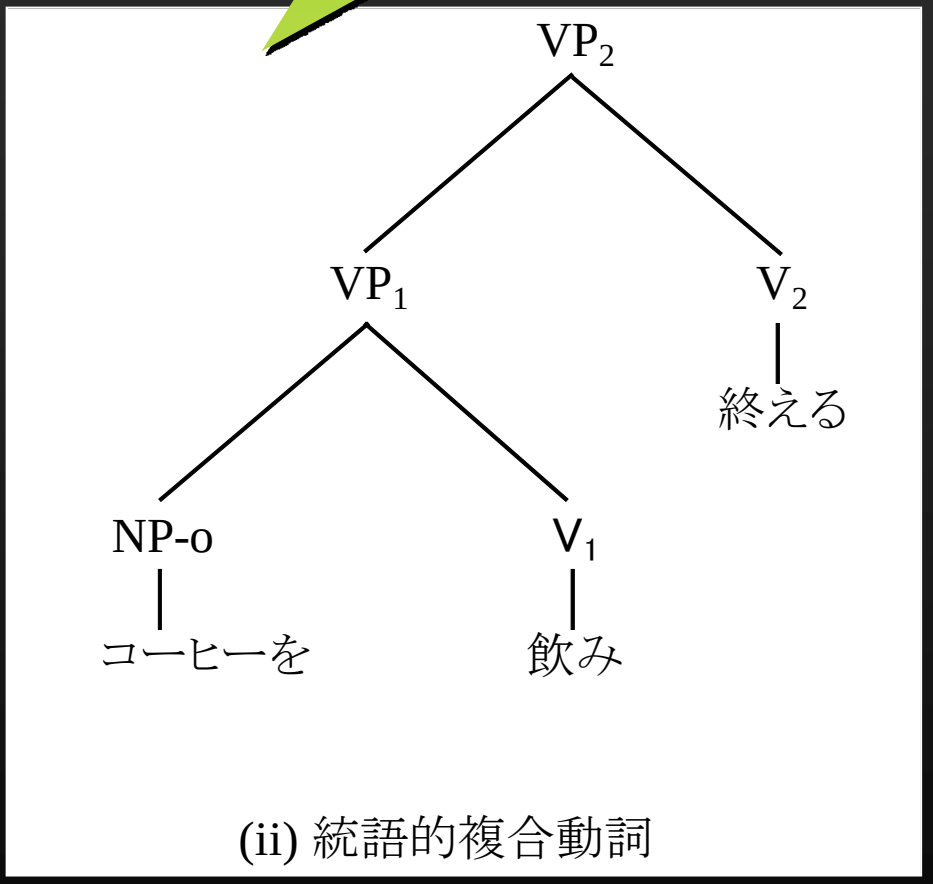
統語的複合動詞

「コーヒーを」という名詞句(NP)が直接「飲む」の動詞(V_1)と結合して動詞句(VP_1)を作る。

この動詞句が後にくる「終える」の動詞(V_2)に結びつき，さらなる動詞句(VP_2)を作るという構造と考える。

二つの動詞は基本的に別々の語であるとみなされる。

[vp2 [vp1 NP-o V_1] V_2]

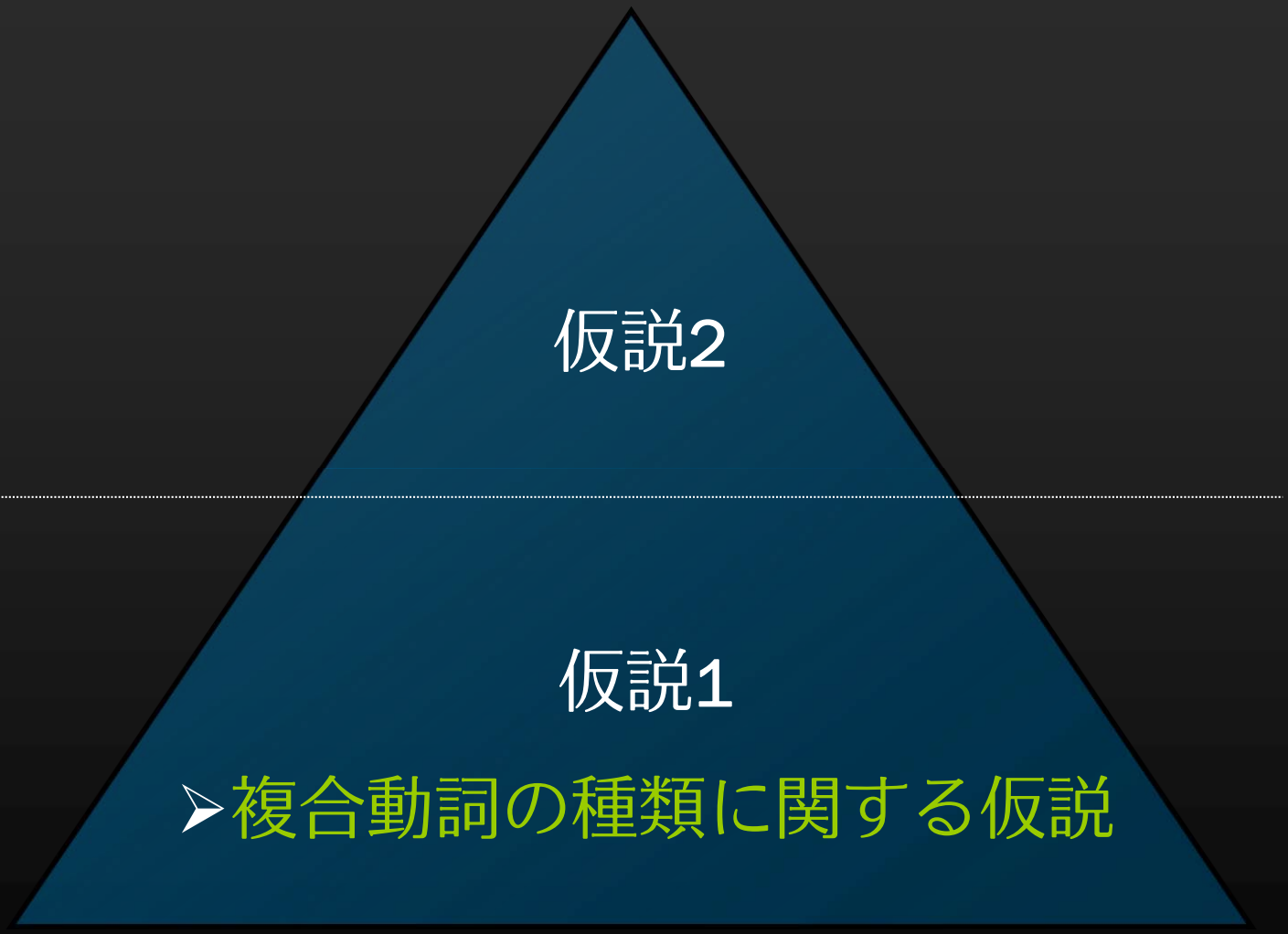


(ii) 統語的複合動詞

複合動詞を新聞と小説のコーパスから抽出

4年間分に限定
小説の約10倍の大きさ

コーパスから得られた語彙的/統語的複合動詞に関する仮説



語彙的/統語的複合動詞に関する仮説１

複合動詞の種類に関して

- 語彙的複合動詞は，「飲み比べる」というように二つの動詞が一つの語彙的な単位を成しており，慣用句のように強く結合していると考えられる。したがって，これら二つの動詞の組み合わせは限定的であり，それほど多様な動詞が組み合わされることはないであろうと予想される。

- 統語的複合動詞は，「終える」(V_2)から分かるように，どのような行為も「終える」ことができるものであれば何にでも結合できるので，多様な第１動詞(V_1)を受けることができると考えられる。

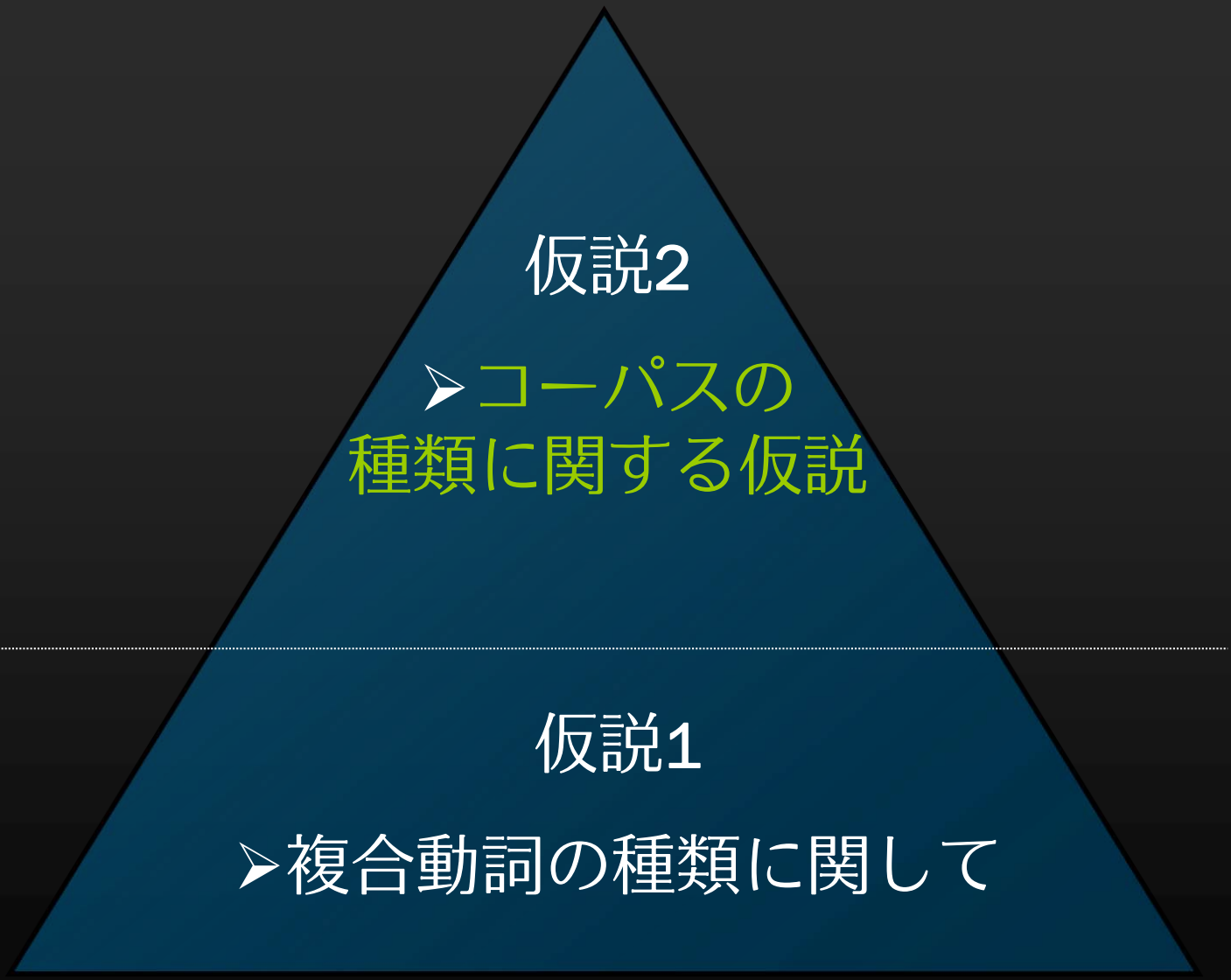
語彙的/統語的複合動詞に関する仮説１

これらの構造の異なる２種類の複合動詞について，通信の数学理論であるエントロピーと冗長度の指標の違いから予測してみる。

仮説1

語彙的複合動詞は，ある特定の二つの動詞の共起パターンは偏っていると考えられるので，
エントロピーは低く，冗長度が高いであろう。
それに対して，統語的複合動詞は，多様な動詞が結合して共起すると考えられるので，
エントロピーが高く，冗長度は低いと予想される。

語彙的/統語的複合動詞に関する仮説2



語彙的/統語的複合動詞に関する仮説 2

コーパスの種類に関して

- 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一定の表現スタイルが採られる。一方，小説は，特定の作家の個性に応じた多様な表現が駆使されている。
- 両者は，書き手の意図と目的において大きく異なっている。

語彙的/統語的複合動詞に関する仮説 2

仮説2

仮に，語彙的/統語的複合動詞が一般的な統語構造を持っているといえるなら，語彙的・統語的複合動詞は，新聞か小説かに関係なく類似した共起パターンを示すことが予想される。

複合動詞の個別性とコーパスの種類との関係

「込む」(V2)

「流れ込む」，「射し込む」，「吹き込む」など

- これらは，水が「流れ込む」，光が「射し込む」，風が「吹き込む」というように，意味的に二つの動詞が一つの動詞として一塊のものとなっていると解釈されるので，**語彙的複合動詞**である。

- 「込む」(V_2)は，ちょうど英語の前置詞の ‘into’のような意味が付加され多様な表現と結びつきやすいので，複合動詞としても多様な動詞(V_1)と結合されやすいと考えられる。

複合動詞の個別性とコーパスの種類との関係

「得る」(V2)

「動かし得る」，「保ち得る」，「考え得る」など

- これらの複合動詞は，「岩を動かし得る」であれば，「岩を動かす」(VP1)ことを「得る」(VP2)という補文構造を成す。「得る」(V2)は，単独でその前の動詞句に結びつく。つまり,これらは**統語的複合動詞**である。
- ちょうど英語の ‘can’や ‘be possible to’のような解釈ができ，「岩を動かす」ことが「できる」または「可能である」という意味になる。

複合動詞の個別性とコーパスの種類との関係

- 第 2 動詞を基準とする複合動詞群について，新聞または小説それぞれのコーパスに特徴的に見られるものが幾つか存在するであろうと思われる。

2つの動詞の共起頻度を他の変数に変換

共起頻度

- 一つずつ増えていく変数であるため，
等間隔の尺度として捉えることができる。

- 数学的な規則にしたがって共起頻度を
他の変数に変換することで，正規性や等質性
を確保することが可能である。

変数の変換

一般的な変換法

平方根変換

- square root transformation

対数変換

- logarithmic transformation

逆数変換

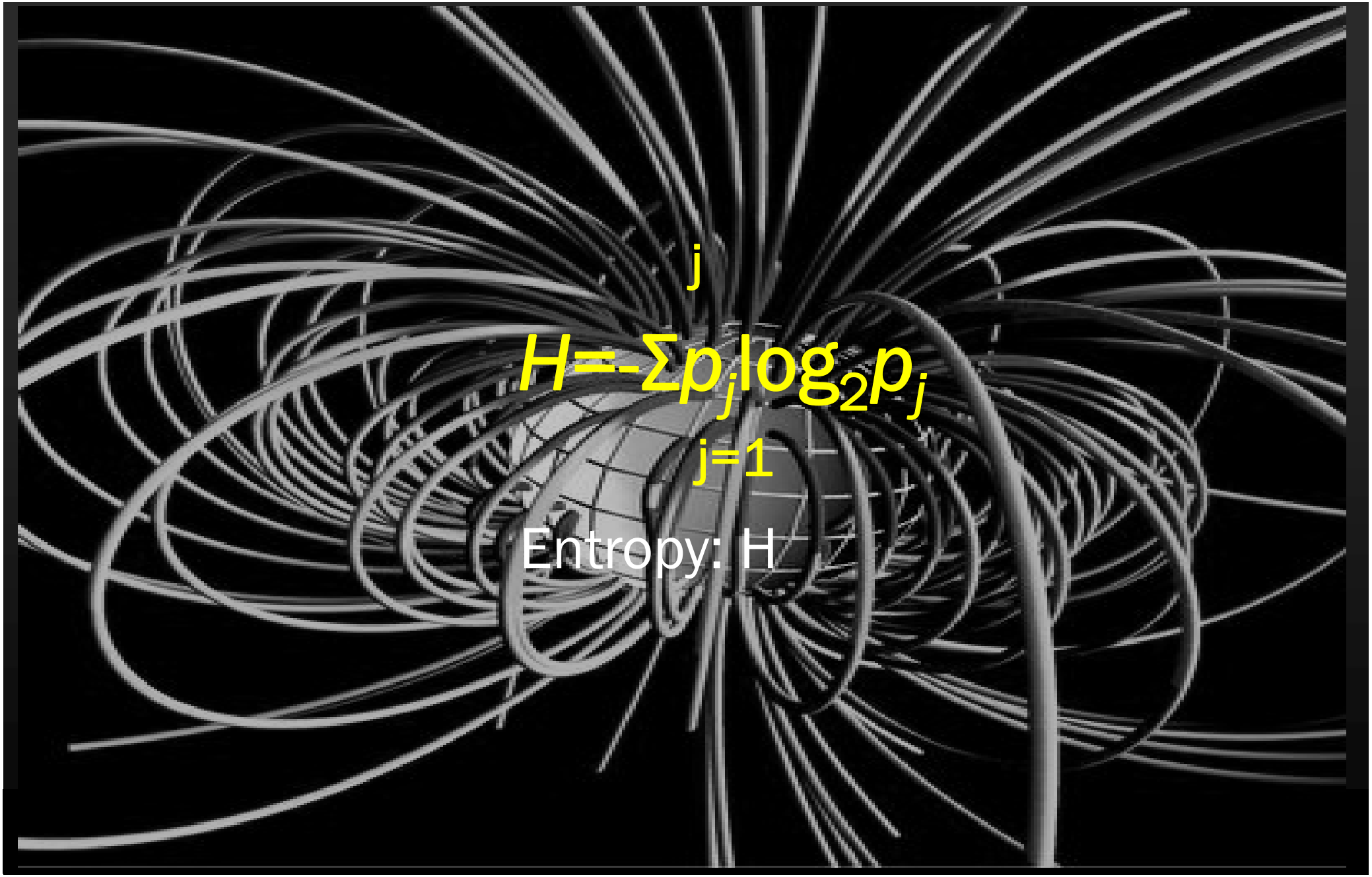
- reciprocal transformation

角変換

- angular transformation

エントロピーと冗長度の指標も，共起頻度の
変数の変換の一種であると考えることができる。

エントロピー (entropy) の公式



複合動詞のエントロピーの計算法

第２番目にくる動詞(V2)を基準とする

例：「歩く（V_2）」を含んだ統語的複合動詞を構成する第１動詞(V_1)の種類

毎日新聞の1991年から1994年までの４年間の記事：
18種類の第１動詞(V_1)と結合

異なり語数

これら18種類の動詞と「歩く」が結合する総頻度:
44回

延べ頻度

売り歩く (7回)　　訪ね歩く (6回)　　・・・・・・

複合動詞のエントロピーの計算法

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

7 (異なり語数)

p_j: 「売り歩く」個別の複合動詞の頻度が
「歩く」(V_2)を基準に作られる
全複合動詞の総頻度に占める割合

÷

44 (延べ頻度)

p_j: 0.159

複合動詞のエントロピーの計算法

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

$\log_2$: 「売り歩く」の場合

$\log_2 0.159$

$\log_2$: -2.652

複合動詞のエントロピーの計算法

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

$p_j \log_2 p_j$: 0.159×-2.652

$p_j \log_2 p_j$: -0.422

複合動詞のエントロピーの計算法

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

-Σ: 同様の計算を18回 (j回) 行い,
すべてを積算して-1をかける。

H: 3.780

冗長度 (redundancy) の公式

$$R=(1-H/Hmax)\times 100\ (\%)$$

Redundancy: R
無駄の程度

- 同じような二つの動詞の組み合わせが
  繰り返し使われる度合い

冗長度 (redundancy) の公式

$$R=(1-H/Hmax)\times 100\ (\%)$$

Redundancy: R
Entropy: H

Hmax: エントロピー最大値

- すべてが等しい確率で生起する場合
- いずれが起こっても不思議ではない混沌としたまったくの無秩序の状態

エントロピー最大値の公式

$$Hmax=log_2 J$$

J: 第 2 動詞を基準とした複合動詞の種類 (異なり頻度)

エントロピー最大値の公式

$$Hmax=log_2 J$$

J: 第 2 動詞を基準とした複合動詞の種類 (異なり頻度)
「歩く」(V2)が作る複合動詞の場合
18種類の複合動詞

異なり頻度

エントロピー最大値の計算法

$$Hmax=log_2J$$

$\log_2 18$

$$Hmax=4.170$$

どの「歩く」(V_2)から作られる複合動詞も同数回だけ出現する，もっとも規則性の無い状態。

冗長度 (redundancy) の計算法

$$R=(1-H/Hmax)\times 100\ (\%)$$

Redundancy: R
Entropy: H

「歩く」のエントロピー: H= 3.780
「歩く」のエントロピー最大値: Hmax= 4.170

1 – 3.780 / 4.170= 0.09348

R= 9.348%

表３　毎日新聞の４年間のコーパスで「歩く」(V_2)が作る複合動詞

#	複合動詞	V_1	共起頻度	p_j	$\log_2 p_j$	$p_j \log_2 p_j$
1	売り歩く	売り	7	0.159	−2.652	−0.422
2	訪ね歩く	訪ね	6	0.136	−2.874	−0.392
3	捜し歩く	捜し，探し	5	0.114	−3.138	−0.357
4	連れ歩く	連れ	5	0.114	−3.138	−0.357
5	さまよい歩く	さまよい	3	0.068	−3.874	−0.264
6	食べ歩く	食べ	3	0.068	−3.874	−0.264
7	泳ぎ歩く	泳ぎ	2	0.045	−4.459	−0.203
8	泊まり歩く	泊まり	2	0.045	−4.459	−0.203
9	めぐり歩く	めぐり	2	0.045	−4.459	−0.203
10	訴え歩く	訴え	1	0.023	−5.459	−0.124
11	踊り歩く	踊り	1	0.023	−5.459	−0.124
12	楽しみ歩く	楽しむ	1	0.023	−5.459	−0.124
13	伝え歩く	伝え	1	0.023	−5.459	−0.124
14	ねり歩く	ねり	1	0.023	−5.459	−0.124
15	触れ歩く	触れ	1	0.023	−5.459	−0.124
16	見歩く	見	1	0.023	−5.459	−0.124
17	呼び歩く	呼び	1	0.023	−5.459	−0.124
18	わたり歩く	わたり	1	0.023	−5.459	−0.124
			44		Σ	−3.780
				エントロピー（H）		3.780
				生産性（J）		18
				累積生産性		44
				エントロピー最大値（Hmax）		4.170
				冗長度(%)		9.348

エクセルで計算すると簡単！

語彙的/統語的複合動詞(V1+V2)の頻度、エントロピー、冗長度

#	V_2の種類	V_2の延べ頻度	V_1の延べ頻度	V_1の重なり頻度	V_1とV_2の延べ頻度	エントロピー	冗長度 (%)
(1) 語彙的複合動詞							
1	込む	295	1,098,690	81	278	5.76	9.10
2	あげる	2,914	45,880	57	174	5.30	9.20
3	切れる	543	64,292	44	119	4.66	14.73
4	取る	5,947	53,493	33	94	4.39	13.04
5	回る	1,021	17,989	27	61	4.27	10.12
6	つく	2,354	8,906	19	45	3.81	10.34
7	歩く	1,554	30,414	18	44	3.78	9.35
8	上がる	1,808	40,283	31	229	3.69	25.56
9	継ぐ	355	20,382	15	33	3.68	5.88
10	死ぬ	1,376	16,929	13	14	3.66	0.97
11	たてる	632	434,024	16	55	3.66	8.46
12	かかる	4,764	61,171	14	25	3.62	4.83
13	替える	135	14,613	15	40	3.58	8.44
14	いれる	2,114	13,410	13	19	3.58	3.35
15	刺す	333	21,434	11	12	3.42	1.19
16	返す	609	9,679	23	45	3.36	25.72
17	出る	7,153	8,513	18	56	3.21	22.94
18	こめる	110	27,164	12	23	3.13	12.63
19	落ちる	755	12,213	11	33	3.07	11.26
20	落とす	599	4,939	11	20	3.05	11.94
21	おろす	253	7,527	10	58	2.88	13.29
22	きる	1,529	136,575	60	496	2.77	53.04
23	入る	6,425	19,260	10	25	2.76	17.06
24	飛ばす	244	1,056	7	13	2.57	8.62
25	つける	460	20,528	9	34	2.51	20.80
26	倒す	140	805	7	14	2.41	14.02
27	殺す	444	1,614	6	14	2.35	8.98
28	起こす	1,392	3,506	6	19	2.07	19.76
29	渡る	573	4,639	7	41	2.00	28.81
30	おりる	431	10,443	7	27	1.68	40.21
31	のぼる	3,417	13,187	4	11	1.68	16.16
32	返る	91	1,487	5	23	1.61	30.68
33	広げる	856	7,812	4	12	1.42	29.09
34	渡す	491	48,068	3	10	1.36	14.13
35	くだる	124	14,816	3	15	1.27	19.69
36	知る	1,993	2,830	5	24	1.14	50.96
37	合わせる	1,109	38,142	15	61	0.88	77.38

#	V_2の種類	V_2の延べ頻度	V_1の延べ頻度	V_1の重なり頻度	V_1とV_2の延べ頻度	エントロピー	冗長度 (%)
(2) 統語的複合動詞							
1	続ける	5,519	539,169	261	1425	6.73	16.21
2	始める	2,983	1,379,861	207	657	6.50	15.55
3	あう	2,302	295,787	170	873	6.16	16.87
4	過ぎる	3,777	368,408	130	515	5.71	18.74
5	まくる	86	708,256	32	66	4.56	8.91
6	終わる	1,884	51,545	31	56	4.50	9.18
7	終える	503	850,402	24	37	4.31	5.90
8	尽くす	687	843,270	26	89	3.72	20.86
9	ぬく	575	724,584	23	131	3.11	31.33
10	かねる	328	1,062,433	18	108	2.82	32.27
11	得る	4,478	362,068	100	1601	0.08	98.77

毎日新聞

88,454,573語から

語彙的/統語的複合動詞(V1+V2)の頻度、エントロピー、冗長度

#	V_2の種類	V_2の延べ頻度	V_1の延べ頻度	V_1の重なり頻度	V_1とV_2の延べ頻度	エントロピー	冗長度 (%)
(1) 語彙的複合動詞							
1	あげる	572	25,037	48	92	5.13	8.17
2	かかる	603	28,586	46	90	5.08	8.01
3	つく	658	35,298	41	107	4.78	10.75
4	たてる	218	7,413	28	57	4.48	6.77
5	出る	1,980	6,355	31	61	4.46	9.95
6	取る	823	4,436	27	45	4.43	6.92
7	きる	311	109,842	36	69	4.42	14.52
8	のぼる	460	33,216	37	142	4.39	15.69
9	回る	162	24,004	29	79	4.31	11.24
10	歩く	368	13,012	23	34	4.26	5.91
11	刺す	150	21,216	20	25	4.21	2.51
12	上がる	182	45,880	62	174	4.19	6.10
13	返す	193	9,719	20	39	4.16	5.28
14	入る	678	20,161	23	55	4.15	8.28
15	殺す	358	48,145	19	36	4.04	4.95
16	落ちる	210	6,852	15	31	3.55	9.24
17	つける	86	20,961	15	30	3.51	10.24
18	合わせる	50	30,796	12	15	3.46	3.58
19	倒す	34	1,603	12	16	3.45	3.69
20	いれる	287	2,702	12	17	3.29	8.31
21	狂う	63	2,581	11	20	3.22	6.87
22	おろす	67	14,421	10	20	3.18	4.15
23	破る	74	3,688	10	14	3.18	4.21
24	くだす	102	21,056	9	10	3.12	1.51
25	渡る	157	8,438	12	30	3.11	13.39
26	返る	67	7,444	9	42	2.62	17.50
27	込む	281	110,771	81	220	2.13	66.48
28	起こす	53	5,335	5	15	1.93	16.87
29	消す	47	706	3	13	1.55	2.30

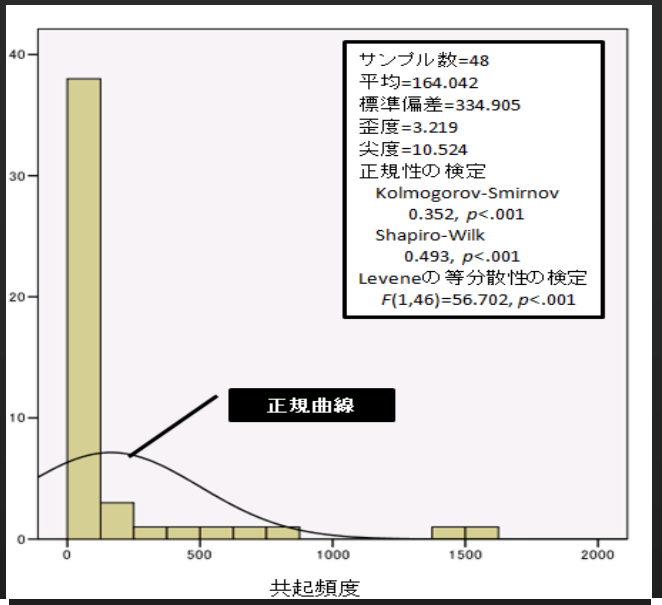
#	V_2の種類	V_2の延べ頻度	V_1の延べ頻度	V_1の重なり頻度	V_1とV_2の延べ頻度	エントロピー	冗長度 (%)
(2) 統語的複合動詞							
1	始める	294	148,945	100	178	6.07	8.67
2	あう	268	133,453	85	173	5.72	10.71
3	過ぎる	617	67,033	66	132	5.44	10.08
4	続ける	146	48,805	50	85	5.28	6.51
5	得る	551	159,594	98	297	5.12	22.56
6	かねる	84	121,923	44	79	5.02	7.99
7	尽くす	20	80,821	10	11	3.28	1.33
8	終わる	64	17,183	12	24	2.98	16.89

青空文庫

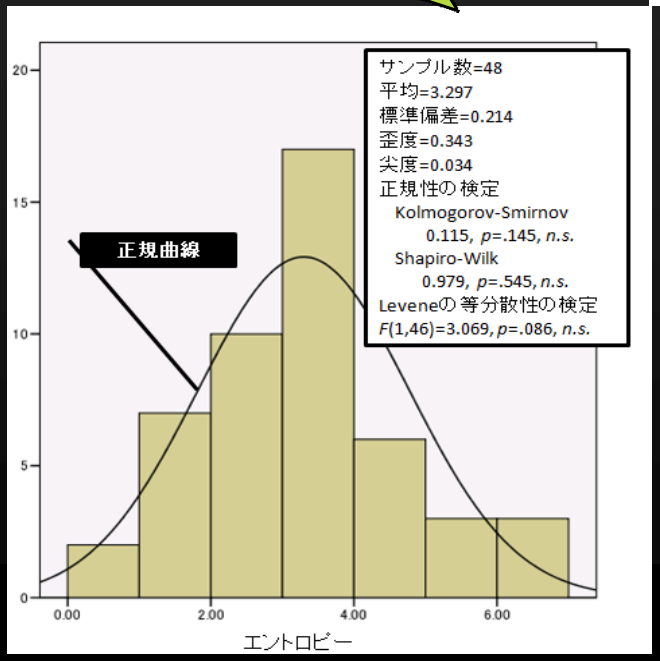
8,370,720語から

分布の比較　ヒストグラム

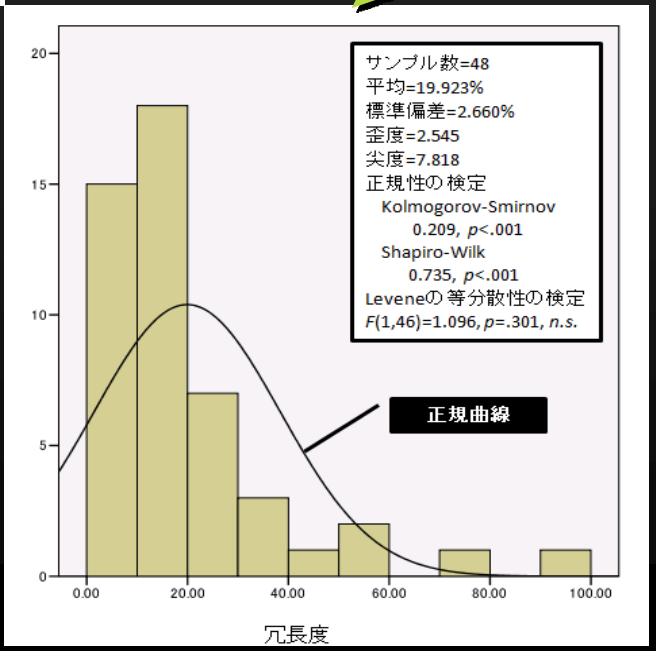
共起頻度



エントロピー



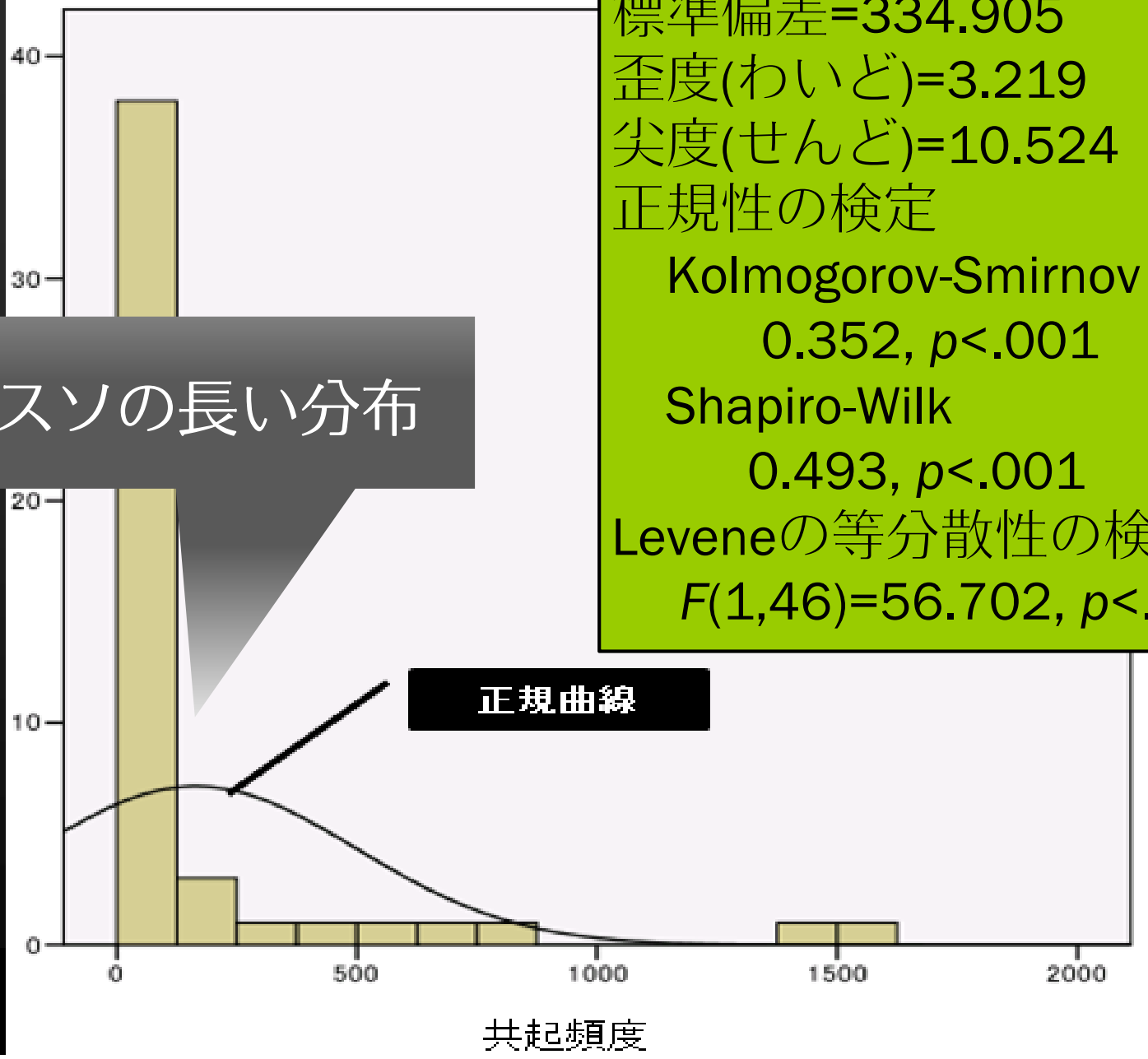
冗長度



共起頻度の分布

サンプル数=48
平均=164.042
標準偏差=334.905
歪度(わいど)=3.219
尖度(せんど)=10.524
正規性の検定
Kolmogorov-Smirnov
0.352, p<.001
Shapiro-Wilk
0.493, p<.001
Leveneの等分散性の検定
$F(1,46)=56.702$, p<.001

非常にスソの長い分布



共起頻度の分布

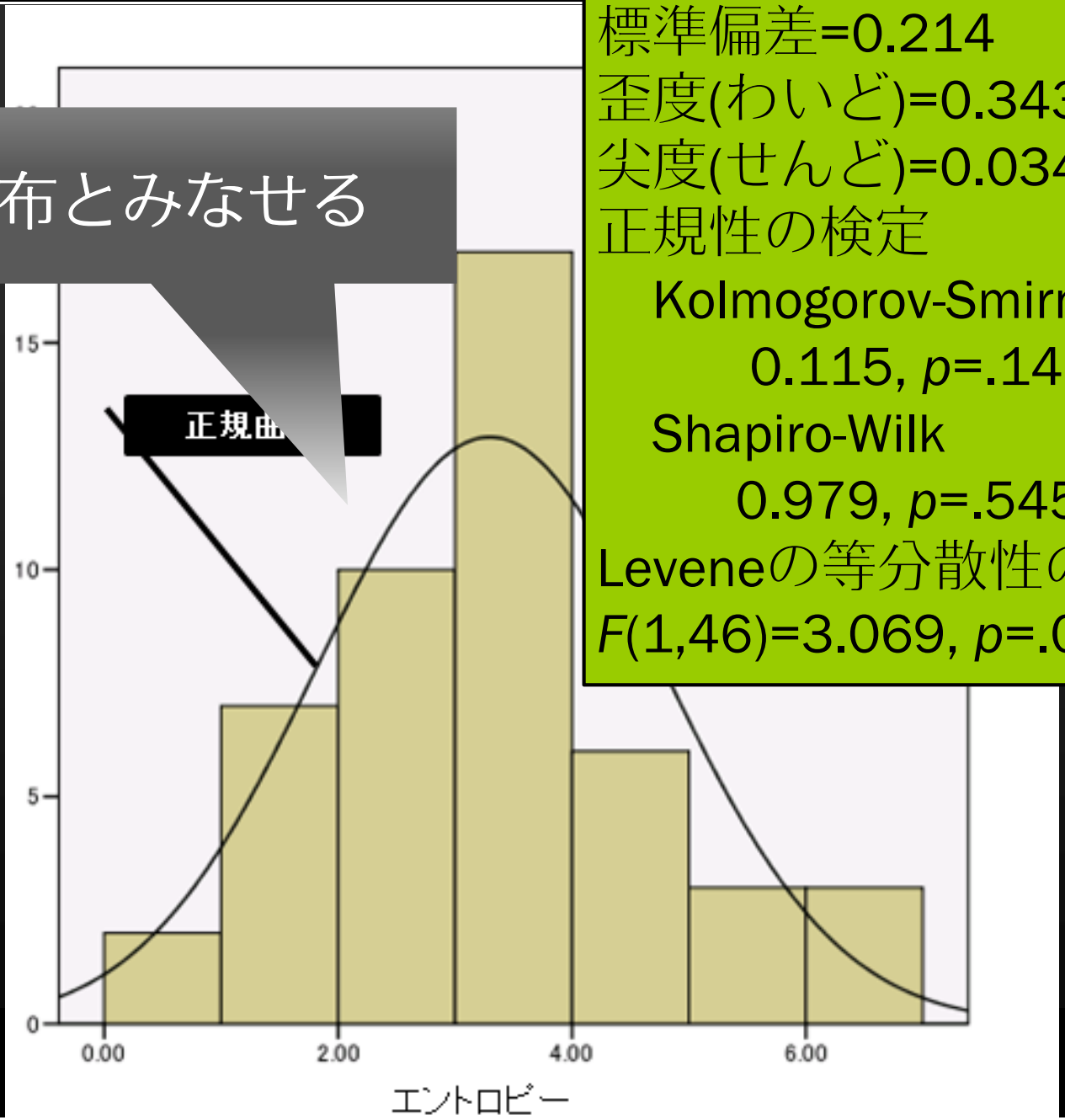
サンプル数=48
平均=164.042
標準偏差=334.905
歪度(わいど)=3.219
尖度(せんど)=10.524
正規性の検定
Kolmogorov-Smirnov
0.352, p<.001
Shapiro-Wilk
0.493, p<.001
Leveneの等分散性の検定
$F(1,46)=56.702$, p<.001

- 正規性: 確保されない
- 等質性: 保証されない

共起頻度の数値に対して
パラメトリック・データの
統計解析を適用することは
できない。

エントロピーの分布

正規分布とみなせる



サンプル数=48
平均=3.297
標準偏差=0.214
歪度(わいど)=0.343
尖度(せんど)=0.034
正規性の検定
Kolmogorov-Smirnov
0.115, p=.145, *n.s.*
Shapiro-Wilk
0.979, p=.545, *n.s.*
Leveneの等分散性の検定
$F(1,46)=3.069$, p=.086, *n.s.*

エントロピーの分布

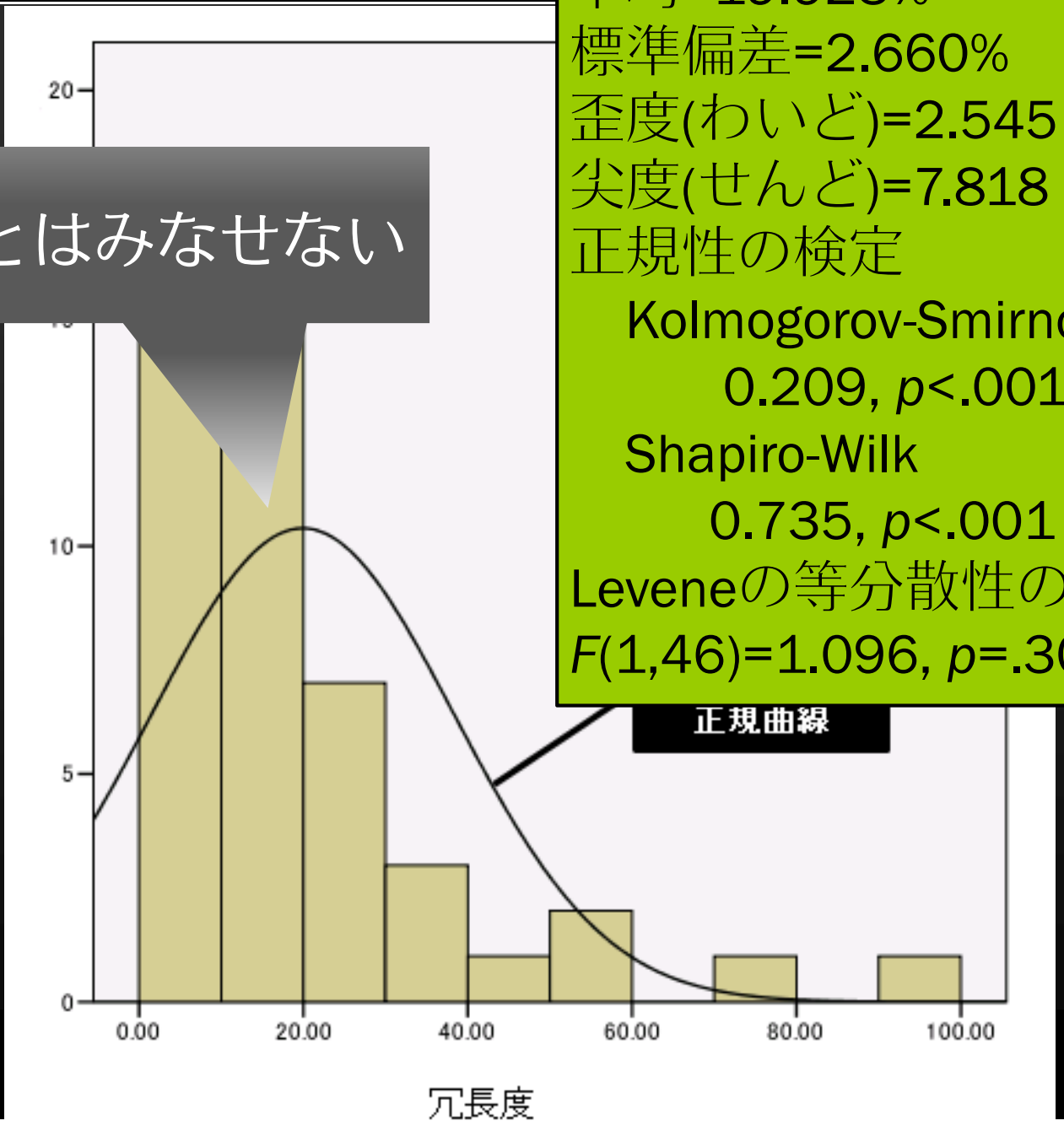
- 正規性: 確保される
- 等質性: 保証される

サンプル数=48
平均=3.297
標準偏差=0.214
歪度(わいど)=0.343
尖度(せんど)=0.034
正規性の検定
Kolmogorov-Smirnov
0.115, p=.145, *n.s.*
Shapiro-Wilk
0.979, p=.545, *n.s.*
Leveneの等分散性の検定
$F(1,46)$=3.069, p=.086, *n.s.*

エントロピーは，きれいな正規分布を示し，パラメトリック・データとしての統計解析ができる。

冗長度の分布

正規分布とはみなせない



サンプル数=48
平均=19.923%
標準偏差=2.660%
歪度(わいど)=2.545
尖度(せんど)=7.818
正規性の検定
Kolmogorov-Smirnov
0.209, p<.001
Shapiro-Wilk
0.735, p<.001
Leveneの等分散性の検定
$F(1,46)=1.096$, $p=.301$, *n.s.*

冗長度の分布

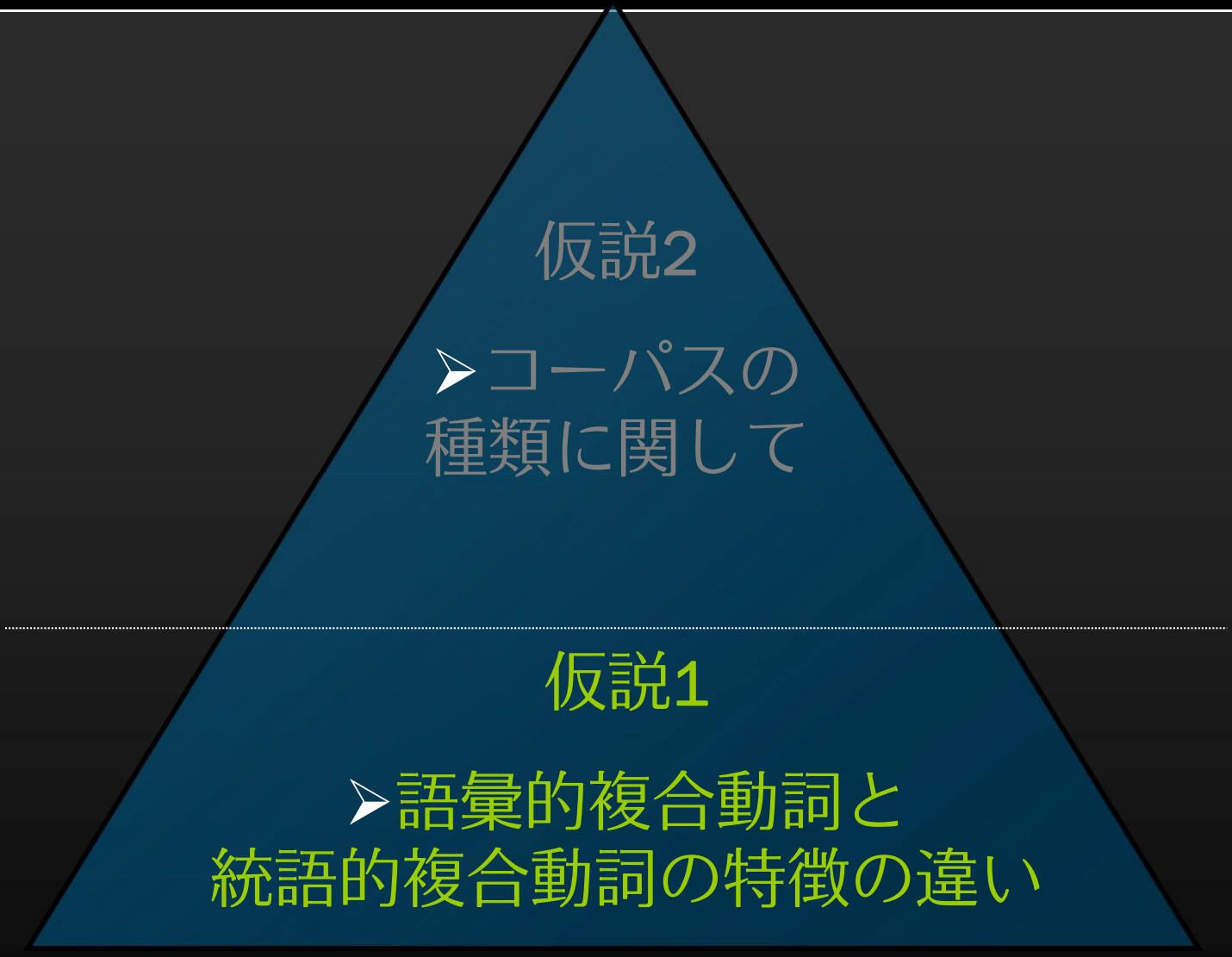
サンプル数=48
平均=19.923%
標準偏差=2.660%
歪度(わいど)=2.545
尖度(せんど)=7.818
正規性の検定
Kolmogorov-Smirnov
0.209, p<.001
Shapiro-Wilk
0.735, p<.001
Leveneの等分散性の検定
$F(1,46)=1.096$, $p=.301$, *n.s.*

*t*検定の適用は問題

分散分析の適用は可能

- 正規性: 確保されない
- 等質性: 保証される

仮説１の検証―語彙的/統語的複合動詞のエントロピーと冗長度による比較



新聞と小説のコーパスとで，語彙的/統語的複合動詞の 2 つのグループについて一元配置の分散分析を行った。

仮説１の検証ー語彙的/統語的複合動詞のエントロピーと冗長度による比較ー新聞

エントロピー

統語的複合動詞
(M=4.138, SD=1.949)
語彙的複合動詞
(M=2.974, SD=1.160)
統語…が有意に高かった
[$F(1,46)=8.946$, $p<.01$]

冗長度

統語的複合動詞
(M=24.963%, SD=25.880%)
語彙的複合動詞
(M=18.425%, SD=15.721%)
有意差はなかった
[$F(1,46)=1.069$, $p=.307$, *n.s.*]

統語的複合動詞のV2が，語彙的複合動詞にV2に比べて多様なV1と結合して複合動詞を作っている。

両複合動詞を作る２つの動詞の結合頻度のパターンの規則性に違いがない。

仮説１の検証ー語彙的/統語的複合動詞のエントロピーと冗長度による比較ー小説

エントロピー

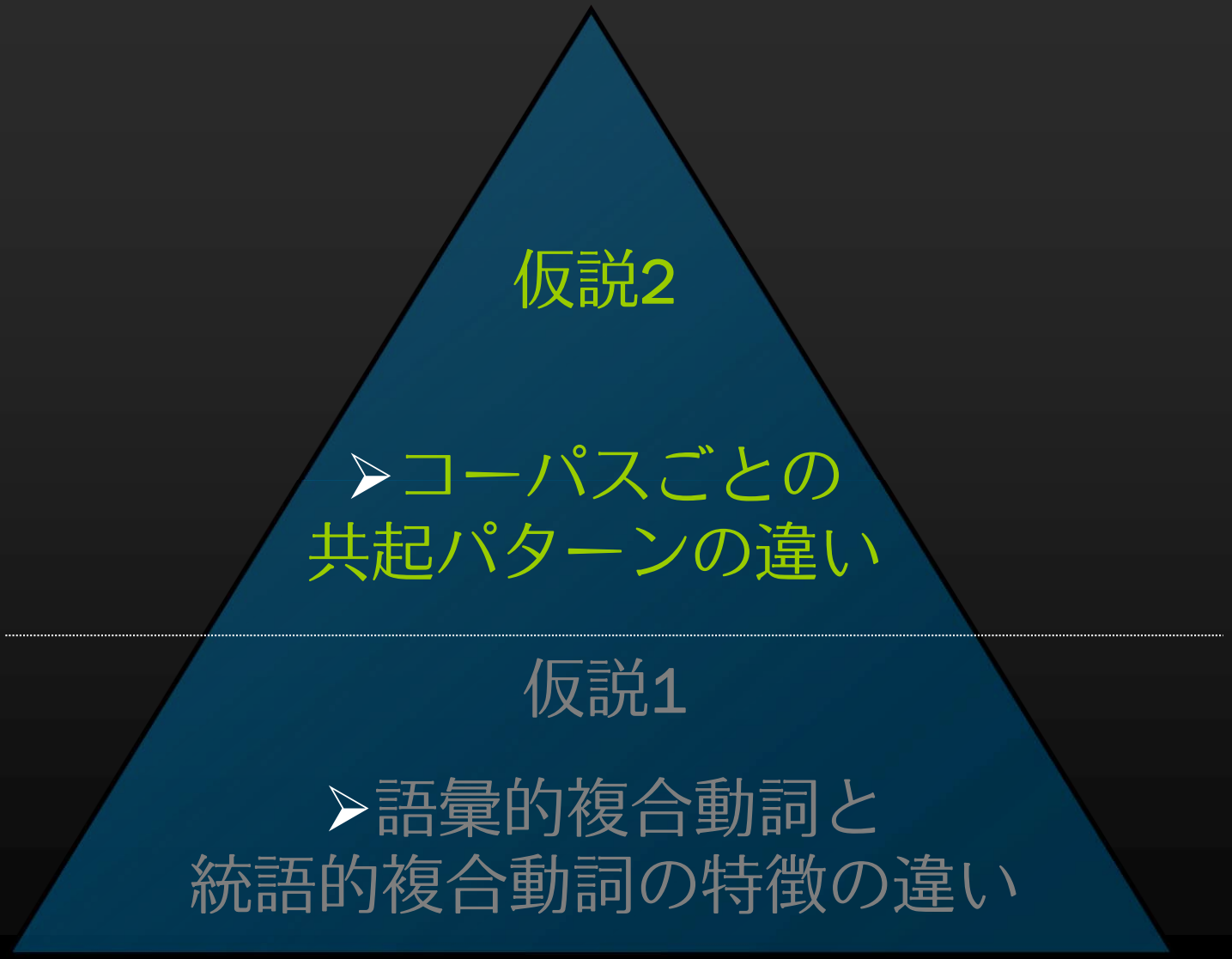
統語的複合動詞
(M=4.864, SD=1.124)
語彙的複合動詞
(M=3.717, SD=0.901)
統語...が有意に高かった
[$F(1,35)=9.144$, $p<.01$]

冗長度

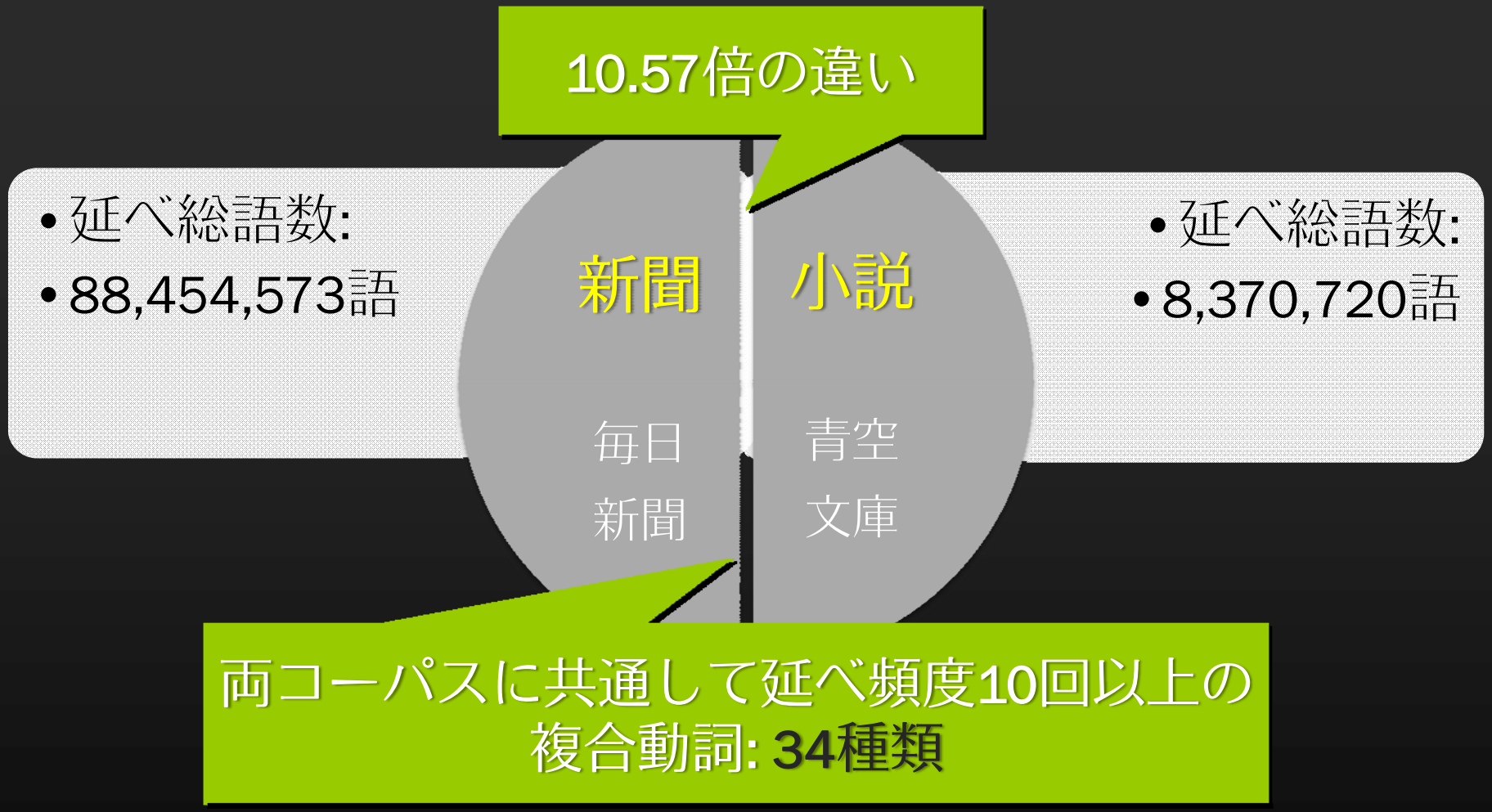
統語的複合動詞
(M=10.593%, SD=6.504%)
語彙的複合動詞
(M=10.117%, SD=11.679%)
有意差はなかった
[$F(1,35)=0.012$, $p=.913$, *n.s.*]

統語的複合動詞のV2が，語彙的複合動詞にV2に比べて多様なV1と結合して複合動詞をつくっている。
エントロピーの指標は，仮説１の結果を支持する。

仮説 2 の検証–語彙的/統語的複合動詞の共起パターンについての新聞と小説との違い



仮説 2 の検証—語彙的/統語的複合動詞の共起パターンについての新聞と小説との違い



複合動詞のエントロピーと冗長度が新聞と小説との間で違うかどうか，反復測定による分散分析で検討した。

仮説１の検証–語彙的/統語的複合動詞の共起パターンについての新聞と小説との違い

エントロピー

新聞
(M=3.422, SD=1.581)
小説
(M=4.080, SD=0.991)
新聞が有意に低かった
[$F(1,33)=6.898$, $p<.05$]

冗長度

新聞
(M=20.958%, SD=19.864%))
小説
(M=10.727%, SD=11.006%)
新聞が有意に高かった
[$F(1,33)=7.358$, $p<.05$]

新聞よりも小説のコーパスの方が，V2を基準としてV1と結合して作られる複合動詞のパターンが，多様性に富んでいる。
新聞のコーパスの複合動詞は，小説のコーパスほどの多様性は無く，より規則的なパターンを示す。

仮説 2 の検証–新聞とコーパスのエントロピーと冗長度による比較

- 新聞は，一般大衆に情報を伝達するために，簡潔で分かりやすい一定の表現スタイルを採っているので，小説よりもエントロピーが小さく，冗長度が大きくなったのであろう。一方，小説は，作家の個性に応じた多様な表現が現れるために，新聞よりもエントロピーが大きく，冗長度が小さくなったのであろう。

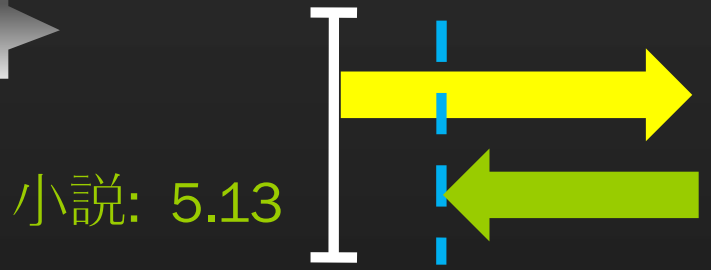
- エントロピーと冗長度は，両者のコーパスの書き手の意図と目的を反映した結果を示しており，両指標がコーパスの違いを比較するのに有効な方法であることが分かる。

エントロピーと冗長度の指標は，仮説 2 を支持する。

個々の複合動詞の記述–新聞と小説それぞれのコーパスに特徴的な語彙的/統語的複合動詞

個々の複合動詞の特徴について，エントロピーと冗長度の差に基づいて記述的に考察する。

「～あげる」のエントロピー

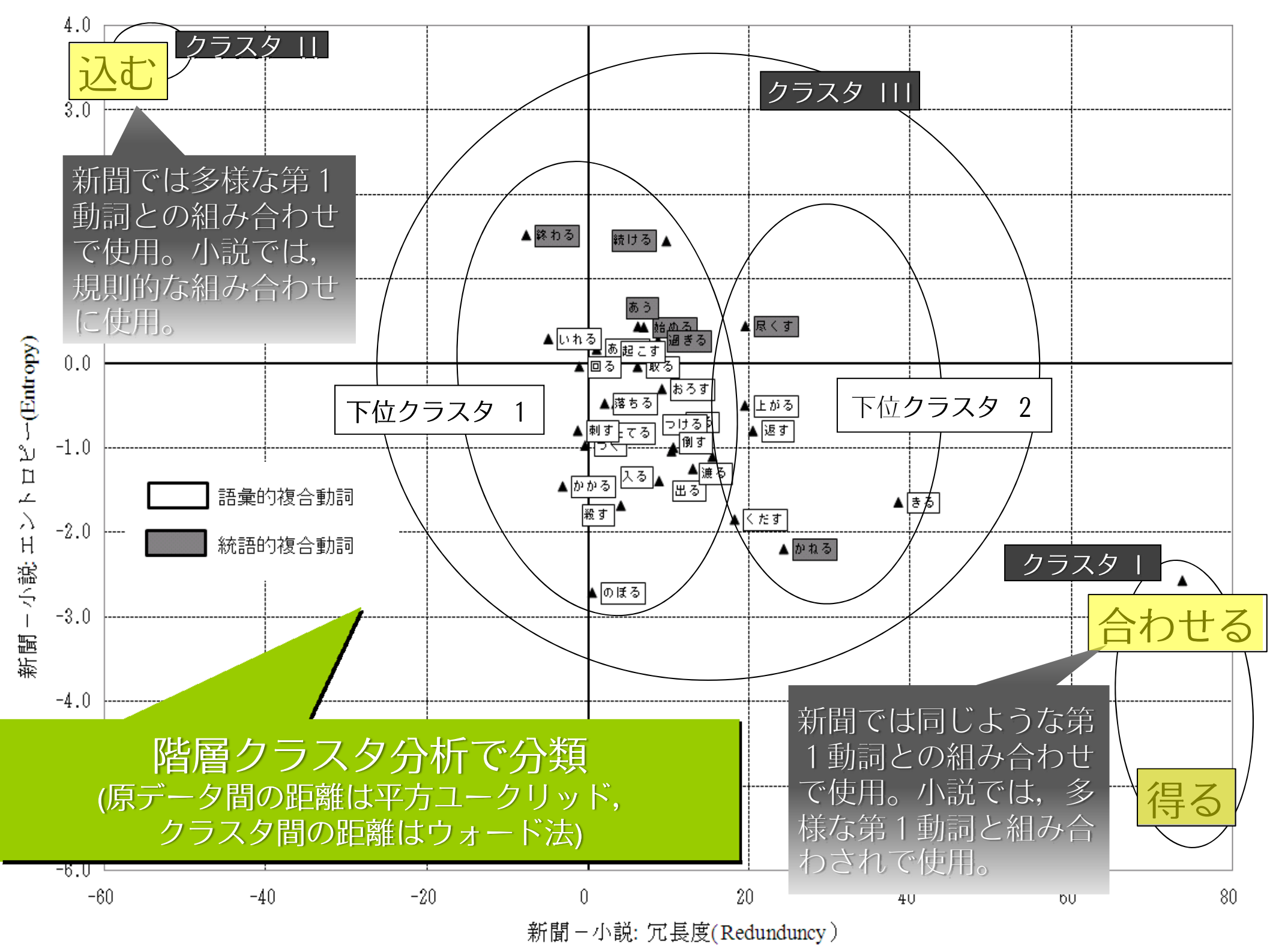


0.17の差

二次元上にプロッティング

「～合わせる」のエントロピー

-2.58 の差


込む
クラスタ II
新聞では多様な第１動詞との組み合わせで使用。小説では，規則的な組み合わせに使用。
クラスタ III
下位クラスタ 1
下位クラスタ 2
クラスタ I
合わせる
得る
新聞では同じような第１動詞との組み合わせで使用。小説では，多様な第１動詞と組み合わされで使用。
語彙的複合動詞
統語的複合動詞
階層クラスタ分析で分類
(原データ間の距離は平方ユークリッド，
クラスタ間の距離はウォード法)
新聞－小説: エントロピー(Entropy)
新聞－小説: 冗長度(Redunduncy)
4.0
3.0
0.0
-1.0
-2.0
-3.0
-4.0
-6.0
-60
-40
-20
0
20
40
60
80
終わる
続ける
あう
始める
過ぎる
いれる
起こす
回る
取る
おろす
落ちる
上がる
返す
刺す
つける
倒す
入る
渡る
かかる
出る
殺す
くだす
かねる
きる
のぼる
尽くす

まとめ

- **変数の変換**：エントロピーと冗長度の両指標をコーパスから得られる共起頻度のパターンの分析に適用し得ることが分かった。特に，エントロピーはきれいな正規分布を示した。
- **2種類の複合動詞の比較**：統語的複合動詞では多様な動詞の結合がみられ，語彙的複合動詞は限定的である。
- **2種類のコーパスの比較**：小説のコーパスの方が，新聞より動詞の結合パターンが多様である。
- **複合動詞の個別性**：新聞または小説に特徴的にみられる複合動詞が存在する。本研究では，新聞と小説での共起パターンにおいて，「込む」，「合わせる」，「得る」の作る複合動詞に顕著な違いがみられた。


Thank you!