Updated: 19 Oct. 2004

英語コーパス学会第 24 回大会ワークショップ
日本大学文理学部
Saturday, 2 October 2004

コーパス言語学のための多変量解析入門

田畑 智司（大阪大学）
tabata@lang.osaka-u.ac.jp

本ワークショップの構成

第 I 部
1. はじめに：コーパス言語学と統計学
2. 多変量解析とは
3. 解析手法概観（対応分析，主成分分析，クラスター分析）
4. 各種法の比較（扱うデータのタイプ，出力など）

第 II 部
5. 多変量解析を行うためのソフトウェア紹介
6. 統計解析言語（環境）R を用いてハンズオンセッション（クリックすると別ファイルへ）
7. 多変量解析に関する文献紹介

1. はじめに：英語コーパス言語学 * と統計学

* 計量言語学・計量文体論・計量文献学などを包含するものとして考える .

Confirmatory approach versus *Exploratory approach*

仮説検証的データ分析
- 理論や仮説を裏付けるデータの検定

仮説探索的データ分析
- 混沌としたデータの中から有用な情報を抽出し仮説構築に役立てる（一種のフィルタリング）

コーパス言語学の文献等で（比較的）よく用いられる統計学的概念や統計値，統計手法：

代表値 (mean, mode, median, etc.)，分散，標準偏差，規準化 (standardisation)，正規分布，*t*-score, MI-score, Log-likelihood 統計量，スピアマンの順位相関，ピアソンの積率相関)

各種検定法・分析法（χ^2 検定，*t* 検定，Mann-Whitney 検定，ANOVA）

統計パッケージとパーソナル・コンピュータが普及し始めた 1980 年代以降多変量解析の活用が進む
→ register variation/text typology, authorship attribution, stylistic studies

Brainerd (1974, 1979, 1980); Bruno (1974); Takefura (1981); Biber (1988, 1993); Biber & Finegan (1989, 1992, 2001); Burrows (1987, 1989, 1992, 1996); Busse (2002); Nakamura (1993, 1994, 1995, 2002); Nakamura & Sinclair (1995); 高橋 (1994); Takahashi (1997); Craig (1999a, b, c, 2001); Opas (1996); Sigley (1997); Tabata (1994, 1995; 2002, 2004)

cf. Bible stylumetry/attribution of disputed authorship

Holmes & Forsyth (1995); Ledger (1995); Linmans (1998); Merriam (1998); Mealand (1999)

2. 多変量解析とは？

大量のデータ（多数の事物や変数）を分類・整理・縮約することでデータの全体像を掴んだり，事物の間にひそむ相互関係や，変数間の相互関係，さらには事物と変数の間の複雑な相互関係を顕在化させるための統計手法の総称。

多数の個体（コーパス言語学で言えば，「テキスト」や「使用域」，「サブコーパスなど」）が多項目の変数（アンケート項目に対する回答や，音素・文字列・語彙・構文など言語項目の生起度等）に関して示す振る舞い（反応）を分析する手法。

対応分析，主成分分析，クラスター分析，因子分析，多次元尺度法，判別分析，重回帰分析，独立成分分析，等々

3. 解析手法概観

3.1 対応分析・コレスポンデンス分析 (Corrrespondence Analysis, CA/ANACOR*/CORRESP)

* *Ana*lyse des *Cor*respondences

（林知己夫の「数量化 III 類」とほぼ同じ原理による反応パターン解析法）

3.1.1 対応分析の特徴（朝野，2000: 27）

・カテゴリーデータ（質的データ）の分析に使うことができる
・多くの変数を小数の次元にまとめることができる
・個体間や変数間の相互関係を視覚化することができる

3.1.2 対応分析の原理（村上 , 1994: 41）

表 1　反応パターン（各作品における色彩語の生起の有無）

作品	A	B	C	D	E	F	G	H
イ			○			○		○
ロ	○			○				
ハ		○		○			○	
ニ			○	○		○		
ホ		○			○		○	
ヘ	○	○						
ト			○	○		○		○
チ		○					○	
リ			○					○

表 2　反応パターンの行と列の並び換え (1)

作品	C	E	D	H	B	G	E	A
ト	○	○	○	○				
ニ	○	○	○					
リ	○			○				
イ	○	○		○				
チ					○	○		
ホ					○	○	○	
ロ			○					○
ハ			○		○	○		
ヘ					○			○

表 1 の行と列の配列を入れ替えて反応パターンの似た作品（同じような色彩語の出てくる作品）と，反応のされ方の似た色彩語を同時に集め，表 2 のように対角線上に最も○印を集中させる。

こうすることで，表 1 に潜んでいたパターンが把握しやすくなる。

しかし，この行列の並び換えは作品数や色彩語数が多数ある場合，パターン行列の複雑度も高いため，1 回の並び換えだけでは（すなわち，1 次元の尺度だけでは）全てのパターンを表しきれない。そのため，2 回目 ,3 回目…（2 次元，3 次元，…）と並び換え行う（次元の拡張を行う）ことになる。表 3 は第 2 次元の尺度で行列の並び換えを行った結果である。

表 3　反応パターンの行と列の並び換え (2)

作品	G	H	C	F	E	B	A	D
ホ	○				○	○		
ト		○	○	○				○
チ	○					○		
リ		○	○					
イ		○	○	○				
ハ	○					○		○
ヘ						○	○	
ロ							○	○
ニ			○	○				○

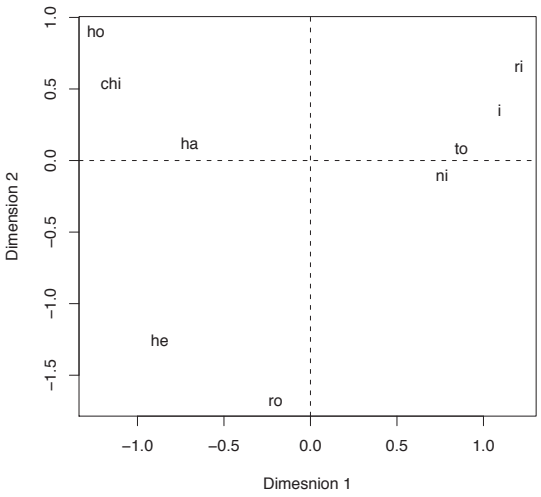

図 1　表 1 のデータの分析結果（第 2 次元まで）：作品

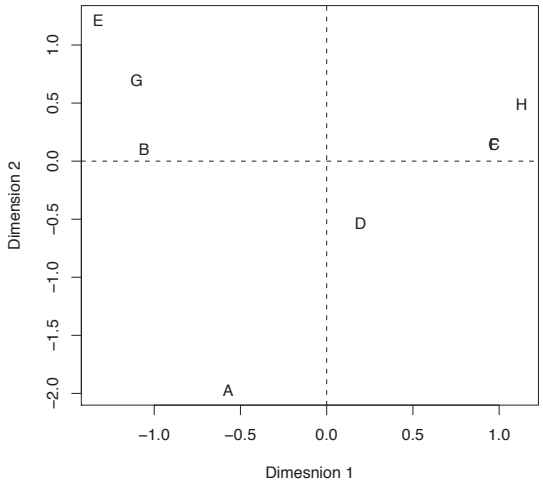

図 2　表 1 のデータの分析結果（第 2 次元まで）：色彩語

表 2 や表 3 のようなデータの行と列の並び換えは，実際には手作業では行うことはできないため，数学的に○印が対角線近くに集まるよう数量を与える操作をする。これは，言い換えるとデータの行と列の相関を最大にすることにほかならない。

3.1.3　対応分析によるデータ分析結果

表 4　相関係数，固有値，寄与率一覧

	1	2	3	4	5	6	7
相関係数(coorelation coefficients)	0.936	0.744	0.581	0.453	0.379	0.379	0
固有値(eigen value)	0.877	0.554	0.337	0.206	0.144	0.144	0
寄与率(proportion accounted for)	41.10%	26%	15.80%	9.60%	6.70%	0.70%	0%

表 5　Row scores（行スコア，サンプル・スコア）

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
イ	1.094	0.350	0.146	0.135	-0.420	-0.144	0.000
ロ	-0.201	-1.682	0.040	0.201	0.601	-0.197	0.000
ハ	-0.699	0.122	-0.307	-0.651	0.311	0.106	0.000
ニ	0.761	-0.106	-0.856	0.100	-0.068	0.064	0.000
ホ	-1.240	0.902	0.035	0.814	0.176	0.006	0.000
ヘ	-0.870	-1.256	0.636	0.143	-0.724	0.165	0.000
ト	0.873	0.084	-0.122	-0.008	0.093	0.068	0.000
チ	-1.153	0.539	0.001	-0.759	-0.348	-0.193	0.000
リ	1.207	0.656	2.082	-0.331	0.575	0.079	0.000

表 6　Column scores（列スコア，カテゴリー・スコア）

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
A	-0.572	-1.974	0.582	0.380	-0.163	-0.134	0.000
B	-1.058	0.103	0.157	-0.250	-0.386	0.174	0.000
C	0.971	0.147	-0.478	0.167	-0.348	-0.031	0.000
D	0.196	-0.531	-0.536	-0.198	0.618	0.086	0.000
E	-1.324	1.212	0.060	1.795	0.464	0.049	0.000
F	0.971	0.147	-0.478	0.167	-0.348	-0.031	0.000
G	-1.101	0.700	-0.156	-0.438	0.122	-0.220	0.000
H	1.130	0.488	1.209	-0.150	0.218	0.010	0.000
I	1.207	0.656	2.082	-0.331	0.575	0.079	0.000

重要！　対応分析の結果を解釈するにあたっては，各次元は互いに直交している（互いに無相関である）ということを常に念頭に置くべきである。

3.1.4 練習問題（データは別ファイル 1. の Biber のデータ参照）

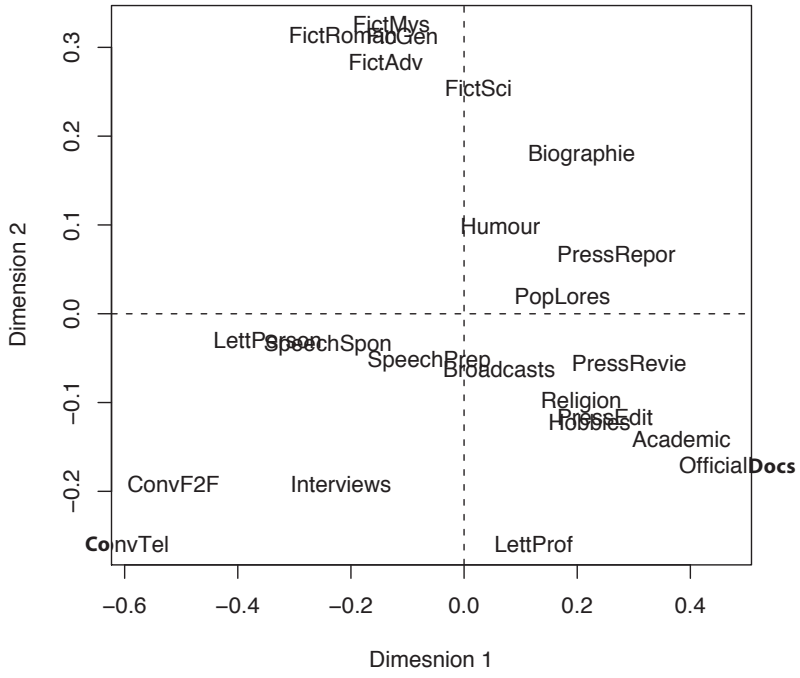

図 3　23 の使用域と言語項目 67 点からなる分割表を対応分析で解析（ サンプル（行）スコア）

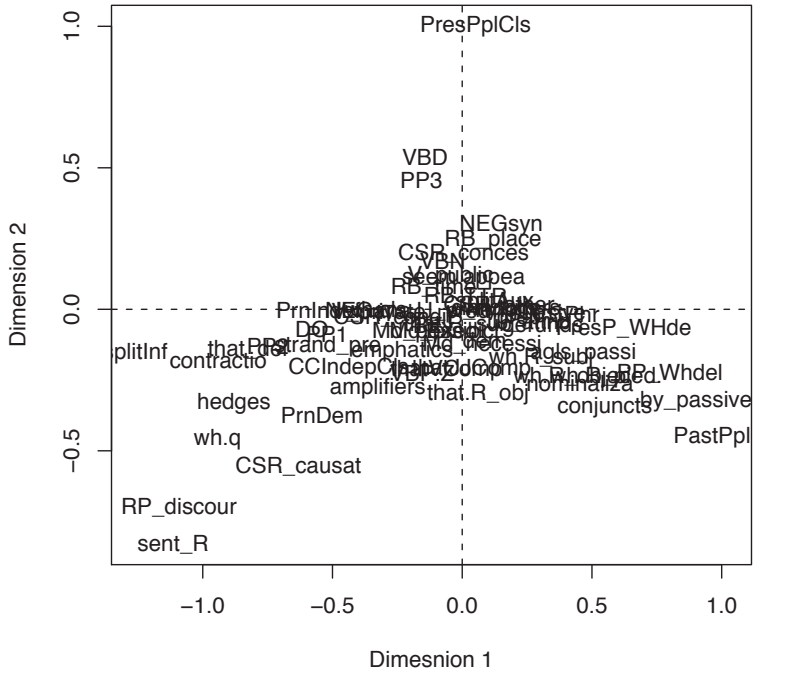

図 4　23 の使用域と言語項目 67 点からなる分割表を対応分析で解析（ カテゴリー（列）スコア）

表 7　Row scores（降順ソート :Dims 1& 2）

Registers	Dim 1
OfficialDocs	0.466
Academic	0.395
PressRevie	0.291
PressRepor	0.269
PressEdit	0.255
Hobbies	0.239
Religion	0.218
Biographie	0.208
PopLores	0.184
LettProf	0.134
Humour	0.086
Broadcasts	0.062
FictSci	0.042
SpeechPrep	-0.062
FicGen	-0.087
FictMys	-0.112
FictAdv	-0.122
FictRoman	-0.208
Interviews	-0.217
SpeechSpon	-0.241
LettPerson	-0.348
ConvF2F	-0.497
ConvTel	-0.581

Registers	Dim 2
FictMys	0.324
FictRoman	0.314
FicGen	0.313
FictAdv	0.284
FictSci	0.254
Biographie	0.178
Humour	0.099
PressRepor	0.065
PopLores	0.017
LettPerson	-0.029
SpeechSpon	-0.035
SpeechPrep	-0.053
PressRevie	-0.055
Broadcasts	-0.063
Religion	-0.100
PressEdit	-0.117
Hobbies	-0.122
Academic	-0.141
OfficialDocs	-0.170
ConvF2F	-0.191
Interviews	-0.192
ConvTel	-0.259
LettProf	-0.259

表 8　Column scores（降順ソート :Dims 1& 2）

Registers	Dim 1
PastPplCls	1.023
by_passive	0.900
PP_Whdel	0.800
PresP_WHde	0.622
wh.R_pied.	0.556
conjuncts	0.550
agls_passi	0.467
nominaliza	0.454
wh.R_objec	0.423
CCPhr	0.400
gerunds	0.312
wh.R_subj	0.304
V_suasive	0.277
JJ_attrib	0.258
prep	0.218
nouns	0.173
dwn.toners	0.168
NEGsyn	0.151

Registers	Dim 2
PresPplCls	1.001
VBD	0.537
PP3	0.457
NEGsyn	0.299
RB_place	0.243
CSR_conces	0.199
VBN	0.170
V_public	0.118
seem.appea	0.103

VBP.Z	-0.162
emphatics	-0.231
CSR_condit	-0.265
IT	-0.295
V_private	-0.321
amplifiers	-0.325
wh.cls	-0.333
NEGana	-0.363
CCIndepCls	-0.437
strand_pre	-0.518
PP1	-0.521
PrnDem	-0.539
PrnIndef	-0.553
DO	-0.579
CSR_causat	-0.631
PP2	-0.749
that_del	-0.827
hedges	-0.879
contractio	-0.940
wh.q	-0.943
RP_discour	-1.092
sent_R	-1.114
splitInf	-1.260

Md_predict	-0.080
Md_possibi	-0.083
PP1	-0.088
dem	-0.108
Md_necessi	-0.124
strand_pre	-0.129
PP2	-0.130
that_del	-0.142
emphatics	-0.147
splitInf	-0.156
agls_passi	-0.156
wh.R_subj	-0.172
contractio	-0.181
CCIndepCls	-0.207
thatJJComp	-0.212
thatVComp	-0.214
VBP.Z	-0.223
PP_Whdel	-0.227
wh.R_objec	-0.239
wh.R_pied.	-0.240
nominaliza	-0.262
amplifiers	-0.279
that.R_obj	-0.299
by_passive	-0.326
hedges	-0.333
conjuncts	-0.345
PrnDem	-0.374
PastPplCls	-0.451
wh.q	-0.458
CSR_causat	-0.554
RP_discour	-0.699
sent_R	-0.829

3.2 主成分分析 (Principal Component Analysis, PCA)

3.2.1 主成分分析の原理

主成分分析も対応分析と同様，多数の変数からなるデータを縮約し，少数の主成分 (PC) でデータの全体像を映し出すテクニックである。解として得られる主成分係数および主成分得点を多次元空間に投影することにより，サンプル（個体）間の関係，変数（カテゴリー）間の関係，さらには，サンプルと変数の関係を視覚的に捉えることが可能である。

主成分分析では，相関係数行列 (correlation matrix) から固有値および固有ベクトルをもとめるか，それとも分散共分散行列 (covariance matrix) から固有値および固有ベクトルをもとめるかというオプションがある。

Burrows (1987)，Burrows & Hassal (1988) では，「テキスト間」の相関行列を基に固有値計算を行い PC を抽出する方法が採られたが，Burrows (1989a) では「変数（語彙）間」の相関行列から固有値計算を行い PC を抽出する方が，テキスト間の言語変異の相をより効果的に視覚化できることが示されており，それ以降 Burrows 自身および彼の方法論を受け継ぐ研究では変数間の相関行列を基礎にする手法を採用している。

> 主成分分析では対応分析に比べてユーザが選べるオプションは多いが，その一方でオプションの選択は解に大きな変化をもたらすことになる。
>
> 主成分分析では（主に）変数の反応のされ方の近似性・異質性を基に分析を行うため，変数間の近似・相異が顕著に表される傾向がある。

表 9　Dickens, *Christmas Books* の 24 人の登場人物の発話における単語 21 語の生起率（発話 1,000 語当たり）

	and	the	you	is	a	of	me	he	was	what	will	that(c)	him	when	so.a.d..	dear	how	would	good	had	never
Scrooge	21.32	31.81	34.25	23.07	25.52	15.38	21.32	6.29	7.69	10.84	12.23	2.80	2.10	1.40	1.05	2.10	1.75	3.15	2.80	2.10	1.05
Toby	29.06	28.36	17.86	28.71	21.01	13.31	10.85	2.45	7.70	7.00	5.95	4.55	1.40	3.15	1.40	3.15	2.45	3.15	4.20	2.80	4.20
MEG	46.16	27.70	22.73	19.18	19.18	12.78	9.94	11.36	5.68	9.23	8.52	7.10	6.39	5.68	9.23	4.97	7.10	6.39	2.13	2.84	3.55
Alderman	25.17	34.40	52.01	24.33	23.49	30.20	6.71	0.00	2.52	10.07	14.26	3.36	1.68	0.00	0.00	0.84	0.84	0.00	4.19	0.00	0.84
Sir Joseph	37.53	45.62	18.40	22.08	23.55	28.70	9.57	12.51	1.47	3.68	11.77	8.83	5.15	2.21	2.21	1.47	1.47	2.94	4.42	0.74	0.00
Will	29.68	25.36	18.89	21.59	28.60	9.17	19.43	5.40	5.40	3.78	7.56	2.70	7.56	10.79	1.62	0.00	3.24	2.16	3.24	1.08	2.16
DOT	39.24	27.25	34.52	15.26	16.72	13.44	13.81	4.72	8.36	5.81	5.45	4.72	5.81	7.63	3.63	10.54	7.27	6.90	2.91	4.36	2.18
John	28.89	25.28	20.95	24.20	23.11	11.92	14.08	14.81	9.39	5.42	7.95	8.67	6.50	2.17	2.17	1.81	5.06	4.33	2.17	6.50	4.33
Caleb	24.17	31.81	25.45	24.17	20.36	11.45	13.36	8.91	7.00	10.18	3.18	3.18	5.09	1.91	2.54	5.09	5.09	6.36	4.45	3.18	5.73
Tackleto	23.00	27.48	48.23	23.56	20.75	19.63	10.10	2.24	4.49	7.85	13.46	4.49	1.12	1.12	1.12	0.56	2.24	1.68	4.49	1.68	0.56
BERTHA	34.92	30.89	31.56	23.51	10.75	11.42	23.51	2.69	4.03	5.37	1.34	4.70	8.06	6.72	13.43	6.04	2.69	5.37	2.69	2.69	8.73
Dr Jeddl	35.24	37.37	21.89	19.75	32.57	26.70	1.60	5.34	4.81	10.14	4.27	5.34	2.14	0.53	1.07	1.07	2.67	2.67	3.74	1.07	3.74
MARION	35.98	14.66	23.32	7.33	8.66	15.32	19.32	9.33	11.33	3.33	4.00	14.66	14.66	5.33	7.99	6.00	2.66	6.66	1.33	7.33	9.99
CLEMENCY	29.63	18.86	30.17	22.09	16.70	15.63	11.31	11.85	17.24	6.47	4.85	7.00	5.93	5.39	1.62	3.23	6.47	6.47	3.77	5.39	2.69
Snitchey	36.41	29.76	26.62	15.27	22.32	23.49	3.52	16.05	9.01	5.09	5.09	7.44	2.35	1.96	1.96	3.92	1.17	4.31	3.13	3.13	0.39
Alfred	48.05	29.00	22.37	19.88	14.91	18.23	12.43	0.00	3.31	7.46	4.97	7.46	0.83	4.97	7.46	7.46	7.46	6.63	3.31	1.66	2.49
Warden	35.64	21.78	25.74	15.18	15.84	21.78	11.88	3.30	3.30	5.28	7.26	5.28	1.98	2.64	1.32	0.00	1.98	4.62	1.32	1.32	2.64
Redlaw	29.71	32.25	22.46	19.57	10.14	27.17	21.38	4.35	4.35	14.13	7.97	5.43	3.62	3.26	2.54	0.00	1.09	3.62	2.17	2.54	1.81
William	33.75	25.74	21.74	36.61	22.31	25.17	2.86	4.58	5.15	16.02	2.29	2.29	5.15	4.00	3.43	2.86	2.29	4.00	2.86	1.14	3.43
MILLY	36.00	24.39	24.00	26.71	15.87	12.78	22.07	17.03	8.13	3.10	5.42	13.94	8.90	6.97	5.03	9.68	5.81	3.10	1.94	3.10	2.71
Philip	37.08	23.17	12.36	22.14	19.57	15.45	10.81	8.24	16.99	4.12	2.57	5.66	7.21	6.18	6.18	1.54	1.03	0.51	5.15	3.09	1.54
MRS TETTERBY	41.58	29.39	33.69	14.34	15.05	20.07	22.94	5.73	15.05	7.89	6.45	7.89	3.58	6.45	8.60	10.75	7.17	2.87	2.15	9.32	4.30
Tetterby	25.48	27.60	36.80	20.52	16.99	21.23	9.20	0.00	7.78	11.32	9.20	7.08	2.83	2.12	0.71	6.37	2.83	1.42	5.66	5.66	1.42
Ghost	34.97	43.90	17.86	21.58	18.60	31.99	8.18	5.21	8.18	3.72	3.72	2.23	3.72	2.98	0.74	1.49	2.23	6.70	1.49	5.95	0.00

表 10　単語 21 語間の相関（ピアソンの積率相関係数）

	and	the	you	is	a	of	me	he	was	what	will	that(c)	him	when	so.a.d..	dear	how	would	good	had	never
and	**1.00**	-0.02	-0.42	-0.35	-0.28	0.00	-0.03	0.16	0.05	-0.30	-0.41	0.37	0.19	0.45	0.62	0.46	0.47	0.42	-0.37	0.13	0.13
the	-0.02	**1.00**	-0.02	0.20	0.38	0.60	-0.26	-0.13	-0.48	0.11	0.24	-0.38	-0.46	-0.42	-0.25	-0.22	-0.27	-0.13	0.17	-0.29	-0.44
you	-0.42	-0.02	**1.00**	-0.06	-0.05	0.13	0.02	-0.38	-0.16	0.27	0.58	-0.16	-0.32	-0.31	-0.15	0.12	0.00	-0.27	0.21	-0.04	-0.14
is	-0.35	0.20	-0.06	**1.00**	0.37	0.01	-0.24	-0.08	-0.22	0.39	0.00	-0.43	-0.24	-0.14	-0.23	-0.25	-0.15	-0.29	0.33	-0.44	-0.20
a	-0.28	0.38	-0.05	0.37	**1.00**	0.13	-0.52	0.08	-0.20	0.11	0.29	-0.43	-0.37	-0.30	-0.56	-0.47	-0.22	-0.43	0.37	-0.46	-0.43
of	0.00	0.60	0.13	0.01	0.13	**1.00**	-0.52	-0.23	-0.31	0.29	0.24	-0.19	-0.46	-0.58	-0.41	-0.37	-0.50	-0.24	0.03	-0.19	-0.50
me	-0.03	-0.26	0.02	-0.24	-0.52	-0.52	**1.00**	0.06	0.17	-0.22	-0.02	0.29	0.43	0.53	0.47	0.36	0.25	0.11	-0.40	0.33	0.40
he	0.16	-0.13	-0.38	-0.08	0.08	-0.23	0.06	**1.00**	0.36	-0.41	-0.15	0.55	0.47	0.12	0.06	0.11	0.18	0.21	-0.25	0.23	0.03
was	0.05	-0.48	-0.16	-0.22	-0.20	-0.31	0.17	0.36	**1.00**	-0.26	-0.37	0.26	0.34	0.32	0.15	0.30	0.27	0.07	0.02	0.71	0.14
what	-0.30	0.11	0.27	0.39	0.11	0.29	-0.22	-0.41	-0.26	**1.00**	0.15	-0.44	-0.43	-0.40	-0.19	-0.13	-0.10	-0.16	0.20	-0.23	-0.08
will	-0.41	0.24	0.58	0.00	0.29	0.24	-0.02	-0.15	-0.37	0.15	**1.00**	-0.08	-0.40	-0.42	-0.45	-0.33	-0.22	-0.50	0.22	-0.28	-0.52
that(c)	0.37	-0.38	-0.16	-0.43	-0.43	-0.19	0.29	0.55	0.26	-0.44	-0.08	**1.00**	0.56	0.19	0.36	0.46	0.28	0.14	-0.27	0.42	0.34
him	0.19	-0.46	-0.32	-0.24	-0.37	-0.46	0.43	0.47	0.34	-0.43	-0.40	0.56	**1.00**	0.58	0.50	0.27	0.15	0.29	-0.34	0.37	0.62
when	0.45	-0.42	-0.31	-0.14	-0.30	-0.58	0.53	0.12	0.32	-0.40	-0.42	0.19	0.58	**1.00**	0.55	0.44	0.47	0.24	-0.26	0.22	0.28
so.a.d..	0.62	-0.25	-0.15	-0.23	-0.56	-0.41	0.47	0.06	0.15	-0.19	-0.45	0.36	0.50	0.55	**1.00**	0.56	0.39	0.34	-0.31	0.26	0.64
dear	0.46	-0.22	0.12	-0.25	-0.47	-0.37	0.36	0.11	0.30	-0.13	-0.33	0.46	0.27	0.44	0.56	**1.00**	0.71	0.36	-0.15	0.51	0.35
how	0.47	-0.27	0.00	-0.15	-0.22	-0.50	0.25	0.18	0.27	-0.10	-0.22	0.28	0.15	0.47	0.39	0.71	**1.00**	0.53	-0.21	0.41	0.23
would	0.42	-0.13	-0.27	-0.29	-0.43	-0.24	0.11	0.21	0.07	-0.16	-0.50	0.14	0.29	0.24	0.34	0.36	0.53	**1.00**	-0.53	0.32	0.40
good	-0.37	0.17	0.21	0.33	0.37	0.03	-0.40	-0.25	0.02	0.20	0.22	-0.27	-0.34	-0.26	-0.31	-0.15	-0.21	-0.53	**1.00**	-0.30	-0.32
had	0.13	-0.29	-0.04	-0.44	-0.46	-0.19	0.33	0.23	0.71	-0.23	-0.28	0.42	0.37	0.22	0.26	0.51	0.41	0.32	-0.30	**1.00**	0.34
never	0.13	-0.44	-0.14	-0.20	-0.43	-0.50	0.40	0.03	0.14	-0.08	-0.52	0.34	0.62	0.28	0.64	0.35	0.23	0.40	-0.32	0.34	**1.00**

3.2.2 主成分の抽出・主成分負荷量／主成分得点の計算

表 10 の相関係数行列を基に固有値計算をおこない，固有ベクトル（主成分係数）を算出する。各固有ベクトルに固有値の平方根をかけることにより，主成分負荷量が求まる→変数間の関係を視覚化。
↓
さらに，固有ベクトルに各変数を標準化 * した元の頻度行列を左から掛けると，主成分得点が得られる。

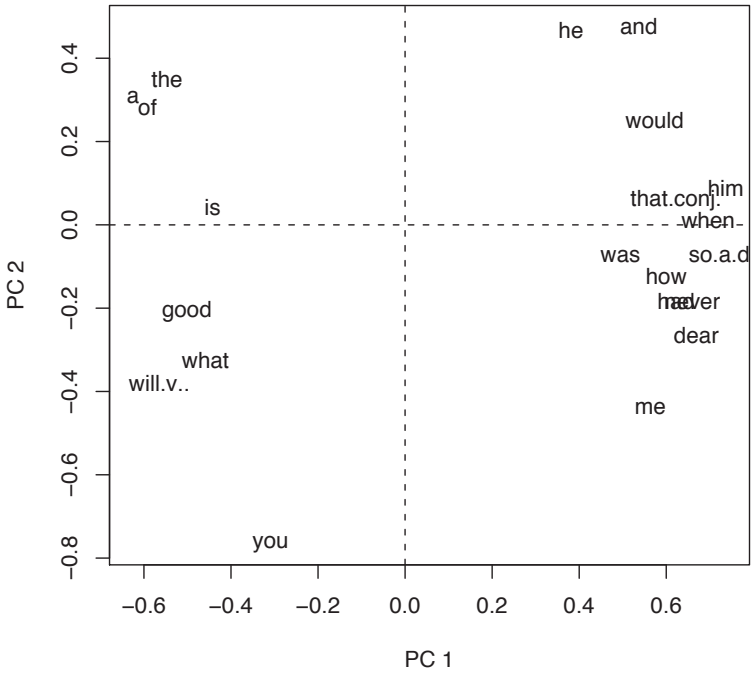

図 4　表 9 の分割表の列間（語彙間）の相関係数行列（表 10）を基に主成分分析を行った結果：主成分負荷量

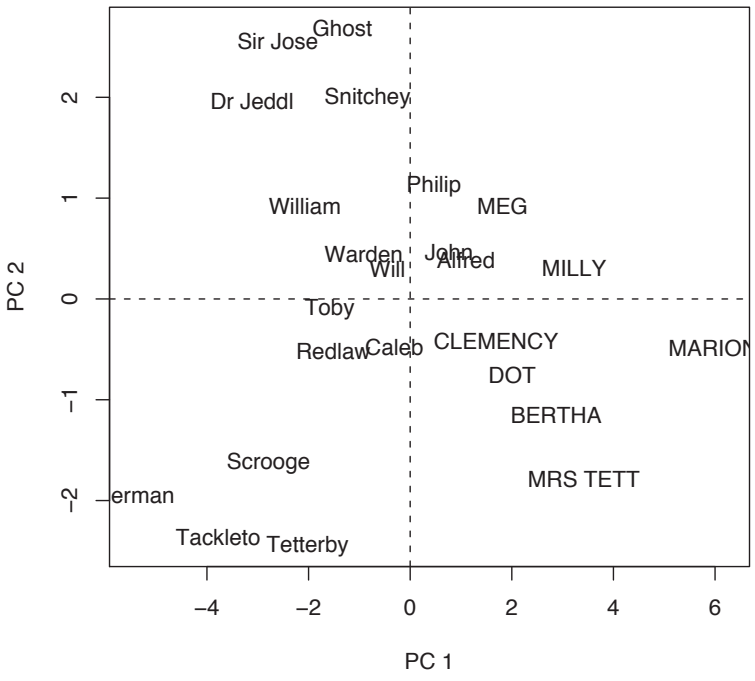

図 5　表 9 の分割表の列間（語彙間）の相関係数行列（表 10）を基に主成分分析を行った結果：主成分得点

表 11　固有値 *

	Eigen values
PC 1	7.02
PC 2	2.05
PC 3	1.95
PC 4	1.78
PC 5	1.49
PC 6	1.30
PC 7	1.01
PC 8	0.93
PC 9	0.78
PC 10	0.65
PC 11	0.48
PC 12	0.41
PC 13	0.33
PC 14	0.32
PC 15	0.25
PC 16	0.11
PC 17	0.07
PC 18	0.03
PC 19	0.03
PC 20	0.01
PC 21	0.00

表 12　寄与率（各主成分によって原データの変動の何 % が説明されているか）

	% accounted for
PC 1	33.44
PC 2	9.78
PC 3	9.26
PC 4	8.48
PC 5	7.08
PC 6	6.17
PC 7	4.80
PC 8	4.45
PC 9	3.71
PC 10	3.11
PC 11	2.27
PC 12	1.94
PC 13	1.55
PC 14	1.52
PC 15	1.18
PC 16	0.54
PC 17	0.33
PC 18	0.17
PC 19	0.15
PC 20	0.05
PC 21	0.01

*　標準化＝平均値が 0，標準偏差が 1 になるように尺度変換すること

重要！ 対応分析の場合と同じく，主成分分析の結果の解釈にあたっては，各主成分は互いに直交している（互いに無相関である）ということを念頭に置くこと。

3.3 クラスター分析 (Cluster Analysis)

3.3.1 クラスター分析とは？

（変数に対する反応が）類似している固体どうしをクラスター（グループ）にまとめ上げていき，固体間の類似関係（質的遠近関係）を視覚化する方法。

3.3.2 クラスター分析の原理 (1)：距離測定の尺度

固体間の類似度をどのようにして計測するか

ユークリッド距離・**標準化ユークリッド距離**

たとえば，下の表 13 で *Kangaroo* と *Lady Chatterley's Lover* の距離を測る場合，各項目の差を二乗したものの総和の平行根をとったものがユークリッド距離ということになる。

しかし，この方法では分散の大きな変数が過大評価されてしまう傾向があるため，各変数を標準化して個体間のユークリッド距離（**標準化ユークリッド距離**）を測定する方法を採用するのが望ましい。

（その他，地理的距離，マハラビノスの距離，ミンコフスキーの距離など）

3.3.3 クラスター分析の原理 (2)：クラスター間の距離の定義の仕方

最長距離法 (complete linkage method)

二つのクラスターから取り出した任意の個体の組み合わせのうち，距離が最大のものをクラスター間の距離と定義する方法。(村上 , 1994: 47)

Ward 法 (Minimum variance method)

二つのクラスターを結合した場合，クラスター内の変動は，結合以前のクラスターの変動の和よりも増加する。この変動の増加分をクラスターの距離と定義する。(村上 , 1994: 47)

（その他，最短距離法，群平均法，重心法，可変法など）

表 13 D. H. Lawrence, Dickens, Smollett の作品 15 点における色彩語 8 語の生起率（100,000 語当たり）

	black	blue	green	grey	purple	red	white	yellow
Kangaroo	85.25	59.94	19.98	39.96	4.00	59.94	95.91	29.97
Lady Chatterley	33.24	34.09	11.08	23.86	2.56	23.86	44.32	31.53
Rainbow	65.61	40.33	17.75	25.28	4.30	36.57	50.55	20.97
Sons & Lovers	70.79	46.57	22.35	30.43	2.48	54.02	79.48	22.35
Women in Love	45.75	35.82	28.66	28.66	7.16	30.31	90.94	23.70
DOT	19.02	6.98	15.85	1.27	0.00	17.12	36.78	3.17
DHT	40.61	5.80	7.74	3.87	1.93	26.11	23.21	0.00
DTTC	27.15	13.21	3.67	6.60	0.73	38.16	20.55	3.67
DGE	41.65	17.85	16.77	7.57	1.08	12.44	29.75	10.28
DOCS	21.15	6.90	20.23	10.57	0.46	17.47	17.93	5.98
SFCF	8.29	1.28	0.00	1.91	0.00	0.64	5.74	1.91
SHC	12.64	4.66	10.64	6.65	1.33	10.64	8.65	7.32
SPP	6.58	2.82	0.94	2.51	0.00	3.13	6.58	0.63
SRR	11.25	6.43	3.75	4.28	0.00	4.82	11.25	2.14
SLG	25.84	8.99	7.86	3.37	0.00	7.86	22.47	1.12

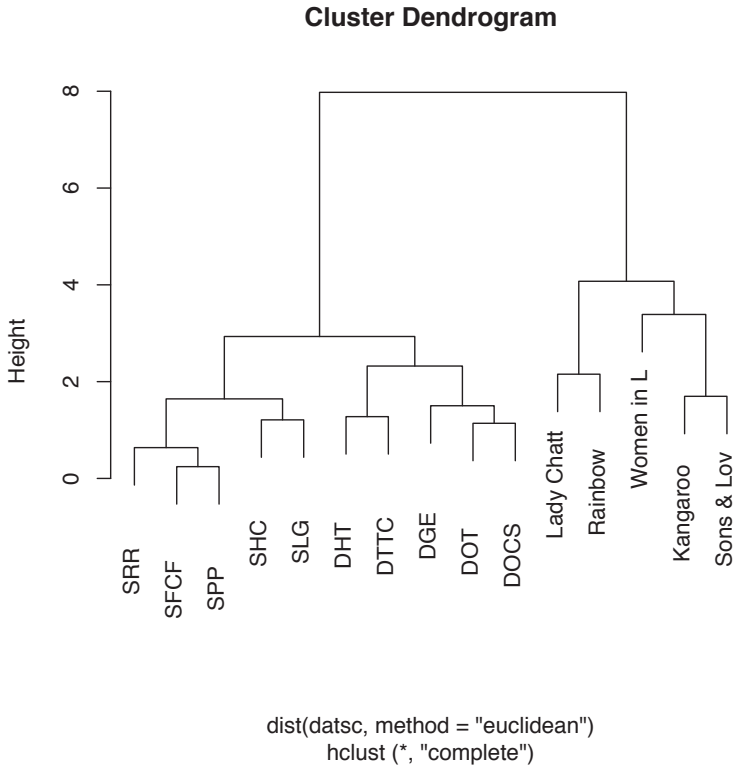

図 6　表 13 の色彩語の生起率分割表を基にクラスター分析を行った結果 (標準化ユークリッド距離）

4. 各種法の比較

・変数のタイプ（質的変数か量的変数か）

量的データの場合は PCA，対応分析，クラスター分析いずれも OK. しかし，質的変数（カテゴリデータ）なら迷わず対応分析。なお, PCA の場合, 変数の尺度が異質なものが混在している場合（例えば,「単語の長さ（文字数）」,「語の頻度」,「センテンス長」など）は相関係数を求め，尺度が同じなら分散共分散行列を選んでみるのもよい。。

・欠損値の有無

欠損値が多い場合は PCA より対応分析を選択。機能語のような欠損値がほとんどない高頻度語データの場合は PCA を選択し，変数間の相関を基に分析を行うとよい。*PCA では個体間の相関を基に分析を行っても言語学的にはあまり興味深い結果が得られないことが経験的に判っているので，変数間の相関を取ること。なお，対応分析ではデータ行列を転置しても全く同じ解が得られる。

・個体の分類か，それとも相互関係の把握か

個体の分類に主たる関心があれば，クラスター分析や判別分析を使うとよい。一方，変数間の関係と個体間の関係を照らし合わせて考察したい場合は対応分析か主成分分析。また，クラスター分析で用いるデータを転置すれば（t(dat) コマンド）変数のデンドログラムを見ることができる。

5. 多変量解析を行うためのソフトウェア（項目をクリックするとウェブサイトへリンク）

統計解析言語（環境）R

SPSS (Win, Mac OS X)
JMP IN (Win, Mac OS X, Linux)
WordStat (http://www.simstat.com/wordstat.htm) (Win)
Statistica (Win)
Le Progiciel R (Mac OS 9)

6. 統計処理言語環境 R を用いて実習 (ここをクリックして別ファイルを参照可能)

あらかじめ表計算ソフト等でタブ区切りのデータを用意しておくこと。

6.1 対応分析

```
>dat <- read.delim(“ファイル名 ”) # タブ区切りデータの読み込み
# パッケージ multiv の読み込み* multiv が利用できない場合別ファイルの 3.2 参照
>library(multiv)

# 行ラベルとして使用する第 1 列を rowvar に代入 ; rowvar の値を行ラベルとして読み込む
>rowvar<-matrix(dat[,1]); rownames(dat)<-rowvar

# 余剰となった第 1 列を削除 ; 列ラベルを colvar に代入
>datca<-as.matrix(dat[,-1]); colvar<-colnames(datca)

# 対応分析 (ca) を実行.
>results<-ca(datca, nf=(min(nrow(datca),ncol(datca))-1))

>ndims<-min(nrow(datca),ncol(datca))-1    # 次元数を ndims に代入.

>rownames(results$rproj)<-rowvar  # サンプル（行）スコアの行ラベルとして rowvar を使用
>rownames(results$cproj)<-colvar  # カテゴリ（列）スコアの行ラベルとして colvar を使用

>results$rproj[,1:ndims]          # 行スコア（個体間の関係）を全て表示
>results$cproj[,1:ndims]          # 列スコア（変数間の関係）を全て表示

# 寄与率 (contribution)【単位はパーセント】を表示
>contribution<-100*results$evals/sum(results$evals)
>print(contribution<-round(contribution, 2))

# 行スコア（個体間の関係）: 第 2 次元までの解を散布図に表示
>plot(results$rproj[,1],results$rproj[,2],type=” n”,xlab=” Dimesnion 1”, ylab=” Dimension 2” )
>text(results$rproj[,1],results$rproj[,2], labels=rownames(results$rproj))
>abline(h=0,lty=” dashed” ); abline(v=0,lty=” dashed” )

# 列スコア（変数間の関係）: 第 2 次元までの解を散布図に表示
>plot(results$cproj[,1],results$cproj[,2],type=” n”,xlab=” Dimesnion 1”, ylab=” Dimension 2” )
>text(results$cproj[,1],results$cproj[,2], labels=rownames(results$cproj))
>abline(h=0,lty=” dashed” ); abline(v=0,lty=” dashed” )
```

6.2 主成分分析

```
>dat <- read.delim(“ファイル名”)     #read.delim コマンドでタブ区切りの頻度分割表を
                                    # 読み込み ,dat に代入 .

# 読み込んだ頻度表の第 1 行（単語ラベル行）は自動的に列ラベルとして
# 認識されるが，第 1 列（人名列）は，自動的には行ラベルと認識されない.
# 行ラベルを設定するために，まず関数 matrix を利用して，第 1 列を rowvar に代入する.
>rowvar <- matrix(dat[,1])

# 関数 rownames を用いて，rowvar を dat の行ラベルに代入.
>rownames(dat) <- rowvar

# 頻度分割表 dat から余剰な第 1 列（人名列）削除し ,datpca として記録.
>datpca <- as.matrix(dat [,-1])

# 関数 colnames を用いて，datpca の行ラベルを colvar として記録.
>colvar<-colnames(datpca)
```

```
# データの行数を nr として記録.
>nr<-nrow(datpca)

# データの行数を nc として記録.
>nc<-ncol(datpca)

#PC 数の上限値（行，列のうち小さい方）
>maxpc<-min(nr,nc)

#PCA のコマンド prcomp を実行し，結果を resultspca に代入する．ここでは scale オプションを
# 有効（=TRUE），すなわち頻度分割表 dat の各列の数値を標準化し，特異値分解によって主成分を抽出する．
>resultspca <- prcomp(datpca,scale=TRUE)

# 主成分負荷量（pcloadings）を表示．
>print(pcloadings<-matrix(resultspca$sdev, nc, maxpc, byrow=T)*resultspca$rotation)

# 主成分得点（pcscores）を表示．
>print(pcscores<-scale(datpca)%*%resultspca$rotation*sqrt(nr/(nr-1)))

# 標準偏差，寄与率，累積寄与率を表示．「標準偏差（standard deviation）」を二乗したもの
#（つまり「分散」）が「固有値」となる．
>summary(resultspca)

# 第二主成分までの主成分負荷量を散布図に表示．
>plot(pcloadings[,1],pcloadings[,2],type=" n",xlab=" PC 1", ylab=" PC 2" )
>text(pcloadings[,1],pcloadings[,2], labels=colvar)
>abline(h=0,lty=" dashed" )
>abline(v=0,lty=" dashed" )

# 第二主成分までの主成分得点を散布図に表示．
>plot(pcscores[,1],pcscores[,2],type=" n",xlab=" PC 1", ylab=" PC 2" )
>text(pcscores[,1],pcscores[,2], labels=rowvar)
>abline(h=0,lty=" dashed" )
>abline(v=0,lty=" dashed" )
```

6.3 クラスター分析

```
# クラスター分析に必要なパッケージ stats を読み込む
>library(stats)

# データの読み込みから整形 # 主成分分析の場合と同じ
>dat <- read.delim( "ファイル名" )
>rowvar <- matrix(dat[,1])
>rownames(dat) <- rowvar
>datclust <- as.matrix(dat [,-1])

# クラスター分析を実施（ここでは，個体間の距離測定に，「標準化ユークリッド距離」，クラスター間の
# 距離の定義には「最長距離法（complete linkage method）」を使うこととする）
#「標準化ユークリッド距離」は「標準化したデータ」の個体間の「ユークリッド距離」を求めたもの。
# 色彩語頻度表を標準化（平均 0, 分散を 1 に変換）するために scale() コマンドを使用している。
# データを標準化することで高頻度の色彩語が過大評価されたり，低頻度語が過小評価されたりするのを防ぐ
# ことができる
>results <- hclust(dist(scale(datclust), method=" euclidean" ))

# 結果を dendrogram（樹状図）として表示
>plot(results)
```

7. Selected References（オンラインのものはクリックすればリンクがたどれる）

多変量解析関連

朝野 熙彦 (2000)『入門 多変量解析の実際』（第 2 版）講談社 .

村上 征勝 (1994)『真贋の科学―計量文献学入門―』（行動計量学シリーズ 6）朝倉書店.

Van de Geer, J. P. (1993a) *Multivariate Analysis of Categorical Data: Applications*. Newbury Park, CA: SAGE Publications.

Van de Geer, J. P. (1993b) *Multivariate Analysis of Categorical Data: Theory*. Newbury Park, CA: SAGE Publications.

柳井 晴夫 (1994)『多変量データ解析法−理論と応用−』（行動計量学シリーズ 8）朝倉書店 .

安本 美典 (1995)『言語の科学―日本語の起源をたずねる―』（行動計量学シリーズ 10）朝倉書店.

統計学一般

青木 繁伸「統計学自習ノート」（online) http://aoki2.si.gunma-u.ac.jp/lecture/index.html

岩淵 千明（編著）(1997)『あなたもできるデータの処理と解析』福村出版.

Oakes, M. (1998) *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh UP.

佐伯　胖・松原　望（編）(2000)『実践としての統計学』東京大学出版会.

Woods, A., P. Fletcher, and A. Hughes (1986) *Statistics in Language Studies*. Cambridge: Cambridge UP.

統計解析言語 R 関連

青木 繁伸「R による統計処理」（online) http://aoki2.si.gunma-u.ac.jp/R/

舟尾 暢男「統計解析 R Tips ―統計解析ソフト R の備忘録 PDF ―」(online) http://cse.naro.affrc.go.jp/takezawa/r-tips.pdf

中澤　港 (2003)『R による統計解析の基礎』ピアソン・エデュケーション.

中澤　港「統計処理ソフトウェア R についての Tips」(online) http://phi.med.gunma-u.ac.jp/swtips/R.html

岡田 昌史（編）(2004)『The R Book ―データ解析環境 R の活用事例集―』九天社.

R-introduction 日本語版 (online) http://buran.u-gakugei.ac.jp/~mori/LEARN/R/R-intro-170.jp.pdf

RjpWiki (online) http://www.okada.jp.org/RWiki/