福山平成大学経営学部紀要
第 18 号（2022），79－99 頁

社会システム分析のための統合化プログラム４２
－テキストＣＲ分析－

福井 正康[*1]・渡辺 清美[*1]

[*1] 福山平成大学経営学部経営学科

要旨：文書に含まれる単語とその語数を用いてコレスポンデンス分析を行い、文書の類似性を調べる手法を、著者らはテキスト CR 分析と呼んでいる。この報告では、分析ソフト College Analysis に組み込んだ、テキスト CR 分析専用のプログラムについて解説する。プログラムは通常のコレスポンデンス分析を実行する部分、その結果を散布図やアニメーションで表示する部分、コレスポンデンス分析の成分の意味を検討する部分に分かれているが、今回は特に最後の成分の意味について実例を用いて考察する。

キーワード：College Analysis、コレスポンデンス分析、文書解析

1. はじめに

文書の出現単語を行、文書名を列として、単語の出現数の 2 次元分割表を作り、コレスポンデンス分析（以後 CR 分析と略す）を用いて、文書を分類する分析が行われることがある。著者らはこれをテキスト CR 分析と呼んでいる。テキスト CR 分析には、通常の CR 分析に比べて以下のような特徴がある。１つは単語の出現数をそのまま使うかどうか、もう１つは出現単語のすべてを使って分析するのか一部を利用するのかである。

これらの問題に対して著者らは参考文献[1]で、一応以下のような結論を得た。前者に対しては文書の長さを変えると単語数も変わり、分析結果も変わることから、単語数は文書ごとにある一定の数に標準化して利用する方がよい。また、後者に対してはある程度安定的な答えが出る必要性から、分割表の中で 0 の占める割合の 0 比率というものを考えて、これが、0.2 程度以下がよいと結論した。また、同じ文献の中で新しい標準化の方法も提案した。

これらの結果を元に、著者らは 2019 年、テキスト CR 分析に特化したプログラムを College Analysis の中に組み込むことにした。このプログラムには、CR 分析の元データとなる単語による文書ごとの単語数の比較表作成機能や、単語数を文書ごとに合わせる標準化機能、統計分析としては新しい、アニメーションによる結果の安定性の確認機能などを加えた[2]。

しかし、アニメーションなどを歴史的な英語の教科書に対して実行すると、組み合わせによっては、分析結果の散布図の形が保たれたまま 1, 2 軸に対して回転するという解釈に

苦しむ結果が得られた。これは CR 分析の軸の意味が変化していることを意味する。これがなぜ起こっているのか、それを知るために、この度再度テキスト CR 分析のプログラムに、軸の解釈を中心とした機能を追加することにした。2019 年のプログラムについては本紀要に未投稿であったため、この論文ではまずプログラムの利用法について復習し、その後成分の解釈を目的とした新しい機能について解説する。

2. プログラムについて

メニュー［分析－多変量解析他－分類手法－テキスト CR 分析］を選択すると図 1 のような分析実行画面が表示される。

図 1　分析実行画面

この画面は、大きく 3 つの部分に分かれている。左上は基本的な分析ツールであり、この部分がテキスト CR 分析の本体である。右側は結果をグラフやアニメーションで表示する部分である。左下は分析結果に現れる成分やグラフの軸について考察を加えるためのデータ解析の部分である。これが今回新しく追加した部分である。この分析実行画面について、次節の単語比較ツールに続いて、順を追って機能別にプログラムの動きを見て行くことにする。

3. 単語比較ツール

テキスト CR 分析では、まず複数の文書から単語の数を取り出し、テキスト間で共通する単語について 1 つにまとめ、すべての文書の語数の合計順に並べ替えるという前処理が必要である。この処理を簡単に行うために、ここではまず以前に作成したツールについて紹介する。

メニュー［ツール－単語比較ツール］を選択するか、2 章図 1 の「単語比較ツールへ」ボタンをクリックすると、図 1 のような「単語比較ツール」実行画面が表示される。

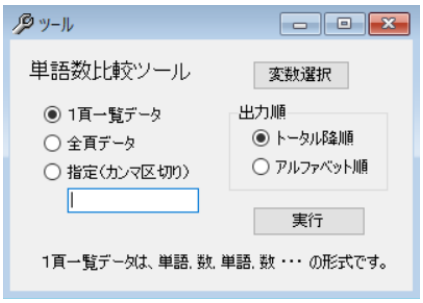
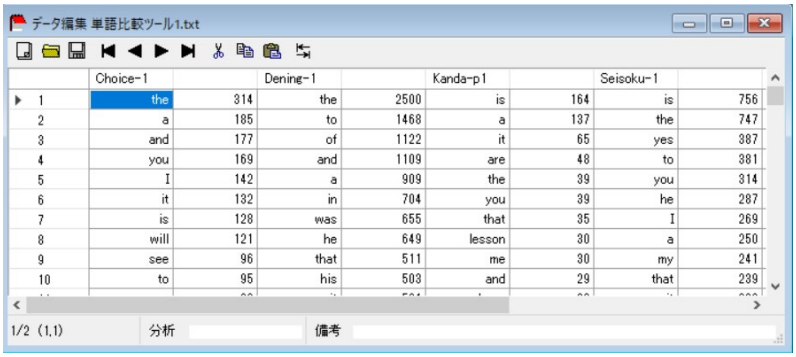


図 1　単語比較ツール実行画面

単語比較のためには、図 2 のように 1 頁に単語とその数、単語とその数、…と並んだデータか、各頁に単語とその数が与えられたデータか、どちらか必要である。単語の並びについては図 2 では文書ごとに降順になっているが、特に指定はない。

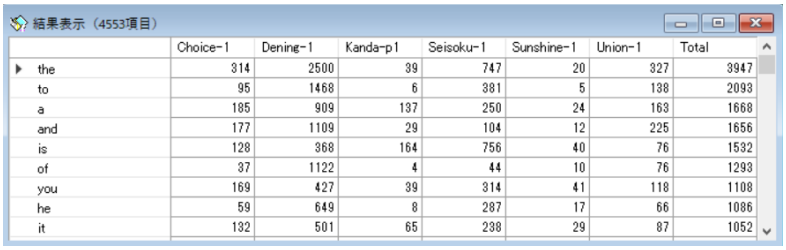


	Choice-1		Dening-1		Kanda-p1		Seisoku-1	
1	the	314	the	2500	is	164	is	756
2	a	185	to	1468	a	137	the	747
3	and	177	of	1122	it	65	yes	387
4	you	169	and	1109	are	48	to	381
5	I	142	a	909	the	39	you	314
6	it	132	in	704	you	39	he	287
7	is	128	was	655	that	35	I	269
8	will	121	he	649	lesson	30	a	250
9	see	96	that	511	me	30	my	241
10	to	95	his	503	and	29	that	239

図 2　単語比較のデータ（単語比較ツール 1.txt）

図 2 で与えられた 1 頁データの場合は、単語比較ツール実行画面の「1 頁一覧データ」を選択し、変数選択で、利用する文書の単語と数の組を指定する。後者の 1 頁 1 文書の場合は、「全ページ」ラジオボタンを選択するか、「指定（カンマ区切り）」ラジオボタンを選択し、利用するデータのページ番号を下のテキストボックスにカンマ区切りで入れておく。

出力は、選択文書全体の語数合計降順の「トータル降順」か「アルファベット順」が選べる。通常、データ形式は「1 頁一覧データ」、出力順は「トータル降順」がよい。この後「実行」ボタンをクリックすると図 3 に示す実行結果が表示される。この結果は単語が頻度順に並べられている。



	Choice-1	Dening-1	Kanda-p1	Seisoku-1	Sunshine-1	Union-1	Total
the	314	2500	39	747	20	327	3947
to	95	1468	6	381	5	138	2093
a	185	909	137	250	24	163	1668
and	177	1109	29	104	12	225	1656
is	128	368	164	756	40	76	1532
of	37	1122	4	44	10	76	1293
you	169	427	39	314	41	118	1108
he	59	649	8	287	17	66	1086
it	132	501	65	238	29	87	1052

図 3　単語比較ツール出力結果

著者らのテキスト CR 分析プログラムは、図 3 の形式のデータを用いるが、単語数の合計を表す「Total」の欄は、分析に不要である。しかし、後に変数選択の中で落とすことができるので、あっても問題はない。このデータは新規に作成されたデータとしても、既存のデータの最後の頁に追加しても、使うことができる。後者の場合は、グリッド出力メニュー［編集－エディタ頁追加］を利用すると便利である。

4. 基本分析ツール

説明を容易にするため、2 章の図 1 分析実行画面の基本分析ツールの部分を切り取って図 1 に再掲する。

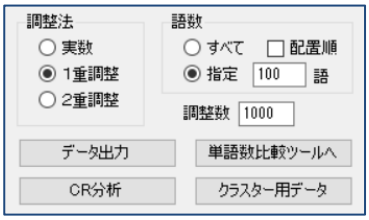
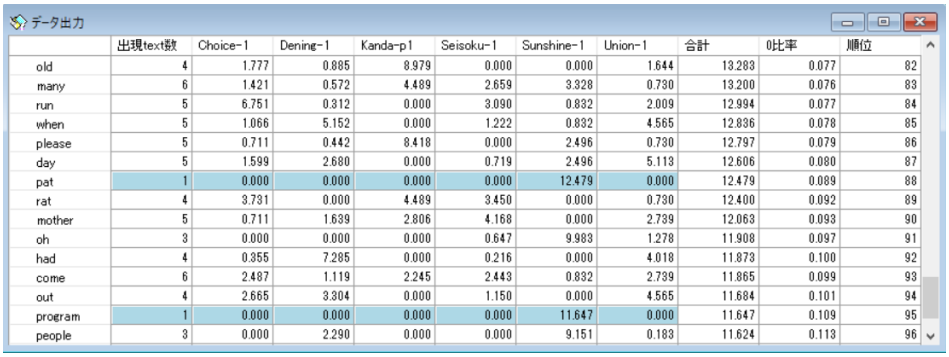


図 1　分析実行画面中の基本分析ツール

テキスト CR 分析では単語数の調整を行うが、このプログラムでは、単語の頻度をそのまま利用する「実数」、単語の頻度をそろえる「１重調整」、単語の頻度をそろえた上で分析に利用する単語数を設定し再度頻度をそろえる「２重調整」の方法を扱うことができる。利用する単語数は「すべて」か、後ろに語数を指定した「指定」を選択できる。このメニューではデフォルトとして、調整法は「１重調整」、語数は「指定」100 語にしている。語数の「調整数」は分析に直接影響を与えないが、「データ出力」の際には値が変わってくるので、見た目が良い程度で記入しておく。デフォルトは 1000 になっている。

「変数選択」で Total を除くすべての変数（文書）を選択し、図 4 の「データ出力」ボタンをクリックすると、図 2 のような出力結果を得る。


データ出力

	出現text数	Choice-1	Dening-1	Kanda-p1	Seisoku-1	Sunshine-1	Union-1	合計	0比率	順位
old	4	1.777	0.885	8.979	0.000	0.000	1.644	13.283	0.077	82
many	6	1.421	0.572	4.489	2.659	3.328	0.730	13.200	0.076	83
run	5	6.751	0.312	0.000	3.090	0.832	2.009	12.994	0.077	84
when	5	1.066	5.152	0.000	1.222	0.832	4.565	12.836	0.078	85
please	5	0.711	0.442	8.418	0.000	2.496	0.730	12.797	0.079	86
day	5	1.599	2.680	0.000	0.719	2.496	5.113	12.606	0.080	87
pat	1	0.000	0.000	0.000	0.000	12.479	0.000	12.479	0.089	88
rat	4	3.731	0.000	4.489	3.450	0.000	0.730	12.400	0.092	89
mother	5	0.711	1.639	2.806	4.168	0.000	2.739	12.063	0.093	90
oh	3	0.000	0.000	0.000	0.647	9.983	1.278	11.908	0.097	91
had	4	0.355	7.285	0.000	0.216	0.000	4.018	11.873	0.100	92
come	6	2.487	1.119	2.245	2.443	0.832	2.739	11.865	0.099	93
out	4	2.665	3.304	0.000	1.150	0.000	4.565	11.684	0.101	94
program	1	0.000	0.000	0.000	0.000	11.647	0.000	11.647	0.109	95
people	3	0.000	2.290	0.000	0.000	9.151	0.183	11.624	0.113	96

図 2　データ出力結果

この結果は一度 1000 語に調整を実行して､その中で頻度の上位から指定語数を選択して表示したものである。これが分析に使うデータである。この中には、参考のために、調整後の単語の合計数や 0 比率などが表示されている。ここでは例として、総頻度が 82 位から

96 位までを表示しているが、この中で水色の網掛けの単語がある。これは 1 つの文書以外では頻度が 0 の単語である。0 比率が低いところの網掛けの単語では、本来利用しない固有名詞などが残っている場合があり、そのような場合にはデータから削除する。データの削除にはエディタのメニュー［ツールー検索］で表示される検索画面で、「行名検索」機能を用いるとよい。

ここで単語の並び順に対して、1 つだけ例外を述べておく。単語を「すべて」選択した場合、「配置順」チェックボックスにチェックを入れると、頻度順ではなく、元の単語の並び順に出力される。これは、特別な単語を入れてその振る舞いを観察する 6 章のデータ解析の際に利用する。

「CR 分析」ボタンをクリックすると、指定された調整法で、指定された語数で CR 分析を実行する。但し、単語数は文書数より多くする必要がある。実行結果を図 3 に示す。

コレスポンデンス分析結果

0比率: 0.112	群	第1成分	第2成分	第3成分	第4成分	第5成分	重み1成分	重み2成分	重み3成分	重み4成分	重み5成分
固有値		0.182	0.146	0.079	0.057	0.016					
相関係数		0.427	0.383	0.281	0.239	0.127					
寄与率		0.379	0.305	0.164	0.119	0.034					
累積寄与率		0.379	0.684	0.848	0.966	1.000					
Choice-1	2	0.161	0.286	0.203	1.787	−1.140	0.069	0.109	0.057	0.427	−0.145
Dening-1	2	1.268	0.731	0.775	−1.513	−0.995	0.541	0.280	0.218	−0.361	−0.126
Kanda-p1	2	−1.856	0.604	0.472	−0.453	0.141	−0.792	0.231	0.133	−0.108	0.018
Seisoku-1	2	0.094	0.052	−2.134	−0.331	0.004	0.040	0.020	−0.600	−0.079	0.001
Sunshine-1	2	−0.019	−2.245	0.449	−0.230	0.031	−0.008	−0.859	0.126	−0.055	0.004
Union-1	2	0.841	0.502	0.442	0.515	2.008	0.359	0.192	0.124	0.123	0.255
the	1	0.908	0.691	−0.112	0.008	−0.132	0.388	0.264	−0.031	0.002	−0.017
is	1	−1.441	0.049	−0.620	−0.631	0.231	−0.615	0.019	−0.174	−0.151	0.029
a	1	−0.945	0.574	0.798	−0.148	0.403	−0.403	0.220	0.224	−0.035	0.051

図 3　CR 分析結果

同じ処理を通常の CR 分析のメニューで実施すると、最初に単語（行名）が表れるようになっているが、ここでは文書の類似性の方が重要であるので、文書名（列名）が最初に並ぶように設定している。表示の項目の意味については、補遺を参照してもらいたいが、特に寄与率と累積寄与率は重要である。

CR 分析の結果を用いてクラスター分析を行い、すべての次元を参照して分類することも可能である。その際、クラスター分析では相関の重み付き成分を利用する方が現実的であるため、「クラスター用データ」ボタンをクリックすると図 3 の四角で囲んだ部分を出力するようにしている。結果を図 4 に示す。

クラスター分析用CRデータ

	重み1成分	重み2成分	重み3成分	重み4成分	重み5成分
Choice-1	0.069	0.109	0.057	0.427	−0.145
Dening-1	0.541	0.280	0.218	−0.361	−0.126
Kanda-p1	−0.792	0.231	0.133	−0.108	0.018
Seisoku-1	0.040	0.020	−0.600	−0.079	0.001
Sunshine-1	−0.008	−0.859	0.126	−0.055	0.004
Union-1	0.359	0.192	0.124	0.123	0.255

図 4　クラスター用データ出力

これをクラスター分析のプログラムのデータとしてデンドログラムを描くことになるが、距離測定法は重み付けをしたことを考慮して、平方ユークリッド距離、クラスター構成法は標準的なウォード法が適していると考える。これらの設定での結果を図 5 に示す。

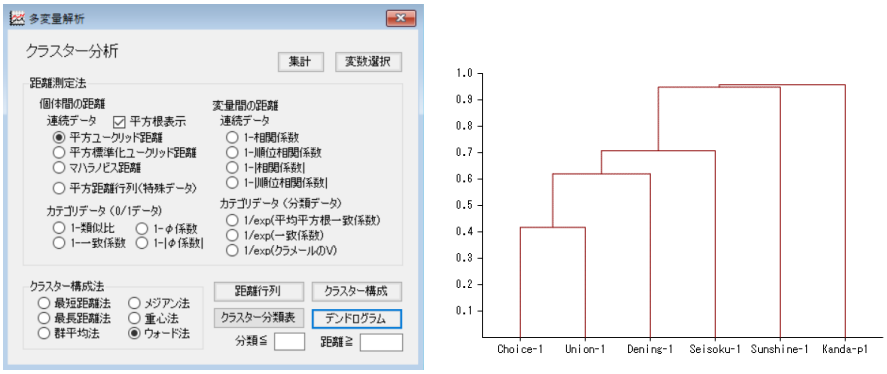


図 5　クラスター分析の実行画面とデンドログラム

5.　グラフ描画とアニメーションツール

次にテキスト CR 分析の結果のグラフ表示を考える。図 1 に分析実行画面のグラフに関する部分を切り取って表示した。

図 1　分析実行画面中のグラフ表示

分析結果を表示するには、「軸設定」ボタンをクリックして、成分を各軸に割り当てる。例えば、x 軸を第 1 成分に、y 軸を第 2 成分にし、「相関重み」を加え、その他の設定をデフォルトの設定にして、「散布図」ボタンをクリックした結果を図 2 に示す。

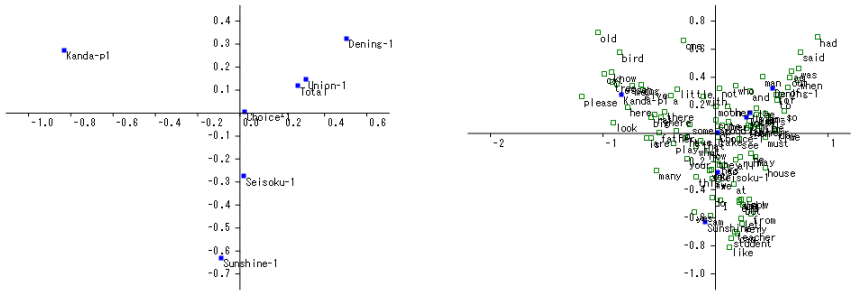


図 2　CR 分析による散布図

左が「列」成分だけの表示、右が「行」成分も含めた表示である。

同様に、「3D」チェックボックスをチェックし、z 軸を第 3 成分にして、その他の設定を図 2 と同じにした散布図を図 3 に示す。但し、分かりにくいのでここでは「列」成分だけにしている。

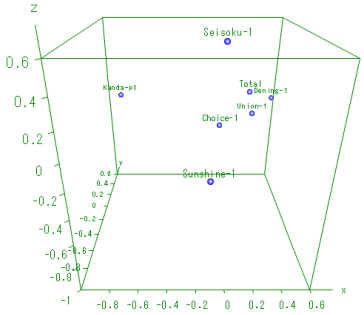
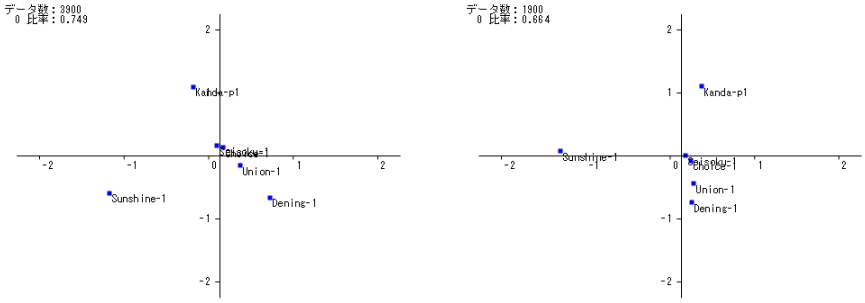


図 3　CR 分析による散布図（３次元表示）

著者らは利用する語数を 100 語に固定してこれまでの計算を行ってきたが、これは 0 比率の値を参考にしながら決めた値である。しかし、語数を決定するとき結果の安定性は重要である。そこで、結果が語数によってどのように変化するかをアニメーションで表示する試みを思い付いた。これは指定された最大語数から、徐々に選択語数を減らして行き、最終的に指定された最小語数まで、散布図が変わって行く様子をアニメーションのように表示する機能である。この動きは紙面上で表現できないが、変化の過程の文書と単語の配置の安定性によって CR 分析の正当性を確認する方法である。

この設定では、単語数の変化を「自動」にするか、「指定」にするか設定できる。「軸」に数値を設定すると絶対値がその数値までの範囲が表示される。図 4 にその過程を簡単に示す。実際に動かしてみると大変興味深いので試してもらいたい。



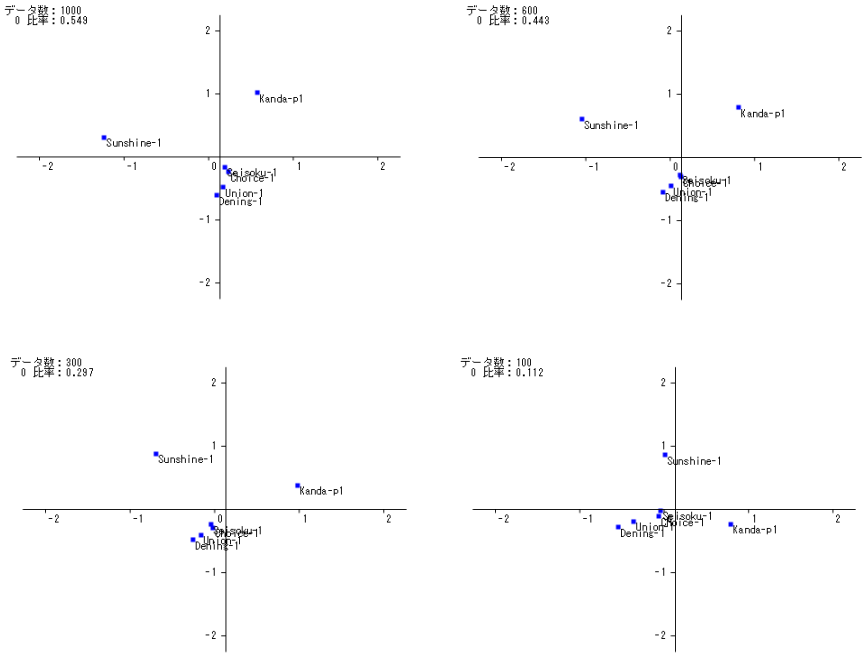


図 4　アニメーション表示の例

6.　データ解析ツールと成分の解釈

CR 分析では成分の意味が明確でない。これは因子分析などと異なる CR 分析の特徴である。特に、テキスト CR 分析では教科書（この章では文書の代わりに教科書を使う）によって単語の数が極端に違う場合があり、この単語の数が教科書の大きな特徴になっている。しかし、この単語数にしてもどの成分が単語数を表しているのか明確ではなく、単語数の似た教科書どうしの比較では、単語数と成分にはあまり関係の見られないこともある。では、これらを調べるには何を見ればよいのか。ここでは定性的な議論であるが、３つの教科書の組についてテキスト CR 分析の特徴を見て行くことにする。

３つの教科書の組としては、１）語数の適度に異なる現代の教科書の組、２）語数の極端に異なる明治期と現代の教科書の組、３）語数の揃った明治期の教科書の組を考える。これらについて、１）ではサンプルの中のテキスト CR 分析 2.txt、２）ではテキスト CR 分析 1.txt (p1)、３）ではテキスト CR 分析 1.txt (p2) を利用する。

分かり易いように、図 1 に分析実行画面からデータ解析ツールの部分を切り抜いた画面を示しておく。

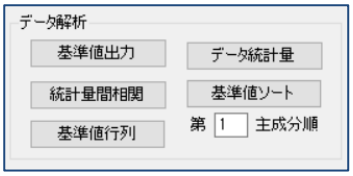


図１　分析実行画面中のデータ解析

１）の場合

テキスト CR 分析では単語をある語数で切り取って分析する。そのため、教科書ごとの頻度が 0 の単語の比率である「文書 0 比率」が重要である。文書 0 比率は単語数の少ない教科書では大きくなる傾向がある。表 1 に実際の結果を示す。切り取る単語数によらず、全単語数との相関係数（網掛け部分）に大きな変動はない。これは単語数が適当に異なる教科書間の興味ある特徴である。

表 1　1 重調整法による文書 0 比率と全単語数との相関係数

語数	50	100	300	500	1000	全単語数
C5	0.140	0.220	0.470	0.548	0.696	1046
C6	0.120	0.210	0.410	0.508	0.678	1085
NH5	0.120	0.290	0.530	0.634	0.767	993
NH6	0.140	0.210	0.407	0.542	0.698	1494
SS5	0.180	0.300	0.453	0.524	0.666	1369
SS6	0.080	0.210	0.403	0.482	0.639	1844
NC1	0.000	0.000	0.053	0.148	0.305	7266
NC2	0.000	0.000	0.050	0.130	0.236	9954
NC3	0.000	0.020	0.083	0.164	0.278	10322
NH1	0.000	0.010	0.043	0.128	0.304	8778
NH2	0.000	0.020	0.067	0.172	0.326	10714
NH3	0.000	0.010	0.107	0.200	0.320	9922
SS1	0.000	0.010	0.083	0.172	0.345	6252
SS2	0.000	0.020	0.120	0.232	0.353	6499
SS3	0.000	0.010	0.070	0.166	0.299	9435
相関係数	-0.912	-0.924	-0.943	-0.943	-0.959	

注）C:Crown, NH:New Horizon, SS:SunShine、数字 5,6 は小学 5,6 年、1,2,3 は中学 1,2,3 年

切り取られたデータから作られた基準値（補遺(A3)式を参照）を $x_{i\lambda}$ とすると、以下のような関係が見られる。

$$\frac{1}{m}\sum_{\lambda=1}^{m} x_{i\lambda} = \frac{1}{m}\sum_{\lambda=1}^{m} n_{i\lambda} \Big/ \sqrt{n_{i\mathrm{g}} n_{\mathrm{g}\lambda}} \;;\; c_m \qquad (1)$$

ここに c_m は教科書の種類 i によらず、切り取った単語数 m だけによる定数である。これは標準化の操作を行ったテキスト CR 分析の特徴かも知れない。著者らはこの指標を「基準値平均」と名付けることにする。

この関係を実際のデータで見てみよう。表 2 に結果を示す。

表 2　基準値平均と語数別標準偏差

語数	50	100	300	500	1000
C5	0.0334	0.0253	0.0131	0.0108	0.0068
C6	0.0359	0.0235	0.0128	0.0098	0.0063
NH5	0.0389	0.0250	0.0122	0.0086	0.0054
NH6	0.0359	0.0228	0.0132	0.0095	0.0060
SS5	0.0322	0.0213	0.0140	0.0113	0.0074
SS6	0.0327	0.0224	0.0138	0.0107	0.0070
NC1	0.0333	0.0233	0.0129	0.0095	0.0066
NC2	0.0310	0.0213	0.0120	0.0091	0.0066
NC3	0.0305	0.0216	0.0119	0.0089	0.0062
NH1	0.0362	0.0239	0.0128	0.0097	0.0063
NH2	0.0323	0.0234	0.0126	0.0094	0.0063
NH3	0.0325	0.0222	0.0123	0.0090	0.0064
SS1	0.0343	0.0231	0.0125	0.0093	0.0065
SS2	0.0327	0.0231	0.0123	0.0091	0.0065
SS3	0.0315	0.0220	0.0120	0.0093	0.0065
標準偏差	0.0023	0.0012	0.0006	0.0008	0.0004

これを見ると教科書による標準偏差は値の 10%以下であり、近似は良い結果を与えている。

次に、$a_{ii}=\sum_{\lambda=1}^{m}x_{i\lambda}^{2}$ で与えられる基準値で作られた基準値行列（補遺(A2)式参照）の対角成分と文書 0 比率の関係を見てみよう。大まかではあるが、以下の関係が見られるようである。

$$(1-\eta_i)a_{ii}=(1-\eta_i)\sum_{\lambda=1}^{m}x_{i\lambda}^{2}\ ;\ d \qquad (2)$$

ここに d は教科書の種類 i にも切り取った単語数 m にもよらない定数である。著者らはこの指標を「対角指標」と名付けることにする。表 3 でこの関係を見てみよう。

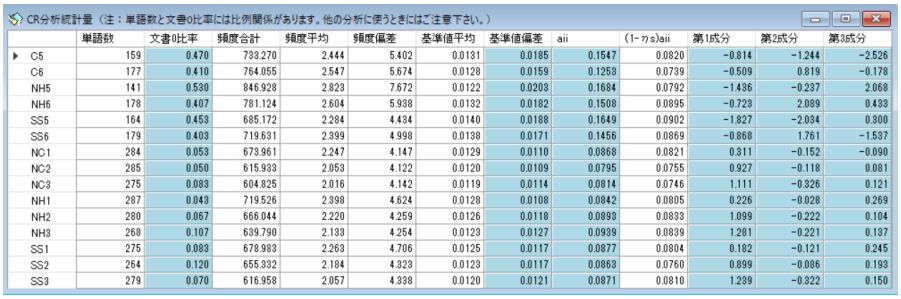
表 3　対角指標

語数	50	100	300	500	1000
C5	0.0776	0.0971	0.0820	0.0944	0.0694
C6	0.0878	0.0824	0.0739	0.0706	0.0549
NH5	0.1258	0.1089	0.0792	0.0645	0.0467
NH6	0.0950	0.0881	0.0895	0.0791	0.0592
SS5	0.0734	0.0692	0.0902	0.0919	0.0758
SS6	0.0817	0.0783	0.0869	0.0891	0.0725
NC1	0.0707	0.0777	0.0821	0.0774	0.0745
NC2	0.0671	0.0701	0.0755	0.0749	0.0769
NC3	0.0693	0.0717	0.0746	0.0727	0.0714
NH1	0.0768	0.0778	0.0805	0.0861	0.0740
NH2	0.0709	0.0776	0.0833	0.0827	0.0764
NH3	0.0765	0.0802	0.0839	0.0790	0.0806
SS1	0.0801	0.0799	0.0804	0.0763	0.0713
SS2	0.0714	0.0755	0.0760	0.0704	0.0705
SS3	0.0750	0.0778	0.0810	0.0841	0.0811

この指標についての全体の平均は 0.0784、標準偏差は 0.0106 である。

次に、これらの指標を含めて、テキスト CR 分析の成分の性質、特に単語数に結び付い

た成分を調べる際に重要と思われる指標について考える。図 2 に分析実行画面の「データ統計量」ボタンをクリックした結果を示す。ここではデータ数を 300 にしている。



CR分析統計量（注：単語数と文書0比率には比例関係があります。他の分析に使うときにはご注意下さい。）

	単語数	文書0比率	頻度合計	頻度平均	頻度偏差	基準値平均	基準値偏差	aii	(1-ηs)aii	第1成分	第2成分	第3成分
C5	159	0.470	733.270	2.444	5.402	0.0131	0.0185	0.1547	0.0820	-0.814	-1.244	-2.526
C6	177	0.410	764.055	2.547	5.674	0.0128	0.0159	0.1253	0.0739	-0.509	0.819	-0.178
NH5	141	0.530	846.928	2.823	7.672	0.0122	0.0203	0.1684	0.0792	-1.436	-0.237	2.068
NH6	178	0.407	781.124	2.604	5.938	0.0132	0.0182	0.1508	0.0895	-0.723	2.089	0.433
SS5	164	0.453	685.172	2.284	4.434	0.0140	0.0188	0.1649	0.0902	-1.827	-2.034	0.300
SS6	179	0.403	719.631	2.399	4.998	0.0138	0.0171	0.1456	0.0869	-0.868	1.761	-1.537
NC1	284	0.053	673.961	2.247	4.147	0.0129	0.0110	0.0868	0.0821	0.311	-0.152	-0.090
NC2	285	0.050	615.933	2.053	4.122	0.0120	0.0109	0.0795	0.0755	0.927	-0.118	0.081
NC3	275	0.083	604.825	2.016	4.142	0.0119	0.0114	0.0814	0.0746	1.111	-0.326	0.121
NH1	287	0.043	719.526	2.398	4.624	0.0128	0.0108	0.0842	0.0805	0.226	-0.028	0.269
NH2	280	0.057	666.044	2.220	4.259	0.0126	0.0118	0.0893	0.0833	1.099	-0.222	0.104
NH3	268	0.107	639.790	2.133	4.254	0.0123	0.0127	0.0939	0.0839	1.281	-0.221	0.137
SS1	275	0.083	678.983	2.263	4.706	0.0125	0.0117	0.0877	0.0804	0.182	-0.121	0.245
SS2	264	0.120	655.332	2.184	4.323	0.0123	0.0117	0.0863	0.0760	0.899	-0.086	0.193
SS3	279	0.070	616.958	2.057	4.338	0.0120	0.0121	0.0871	0.0810	1.239	-0.322	0.150

図 2　「データ統計量」実行結果

これには開発者が重要であると考える指標が教科書ごとに並んでいるが、教科書ごとの文書 0 比率は単語数と関係のある重要な指標であろう。また、基準値から作られる基準値行列 a_{ij} は、固有方程式を与えることから重要な要素であるが、特に対角成分 a_{ii} は各教科書のデータのばらつきを与えるものである。またこの指標は(2)式から文書 0 比率と関係しているとも考えられる。同様にして、教科書ごとの基準値の標準偏差も意味を持つかも知れない。これに、各教科書の固有ベクトル成分を 3 つまで加え、検討すべき指標と考えた。これらの指標については、青色に網掛がされており、簡単に教科書ごとの相関を見ることができるようになっている。

これに対して、上で述べた基準値平均や対角指標は、あまり教科書による変動が期待されないので、確認をするためのデータである。また、基準値の元となる頻度については、直接固有方程式の行列を与えるものではないので、網掛けが行われていない。もちろん相関を求めることが必要な場合は、図 2 のデータをグリッドエディタにそのままコピーし、相関を調べることもできる。

次に、先に述べた網掛けの指標の相関を求めてみよう。図 1 のメニューの中の「統計量間相関」ボタンをクリックすると、図 3 のような主要統計量間の相関行列が得られる。

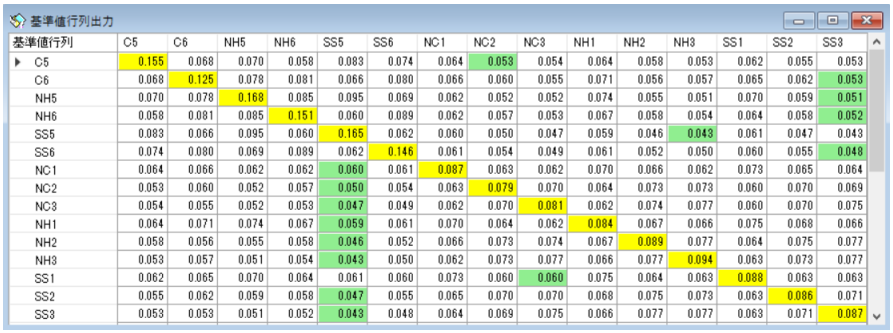
主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.984	0.977	-0.898	0.088	-0.123
基準値偏差	0.984	1.000	0.994	-0.888	0.053	-0.066
aii	0.977	0.994	1.000	-0.922	0.037	-0.091
第1成分	-0.898	-0.888	-0.922	1.000	0.006	0.042
第2成分	0.088	0.053	0.037	0.006	1.000	-0.007
第3成分	-0.123	-0.066	-0.091	0.042	-0.007	1.000

図 3　主要統計量間の相関行列

ここでは文書 0 比率と第 1 成分とが強い相関を持っているので、第 1 成分が単語数を通じた難易度を表しているものと解釈できる。

テキスト CR 分析の固有方程式の行列を与える基準値行列 a_{ij} については、図 1 のメニューで「基準値行列」ボタンをクリックすると、図 4 のように与えられる。



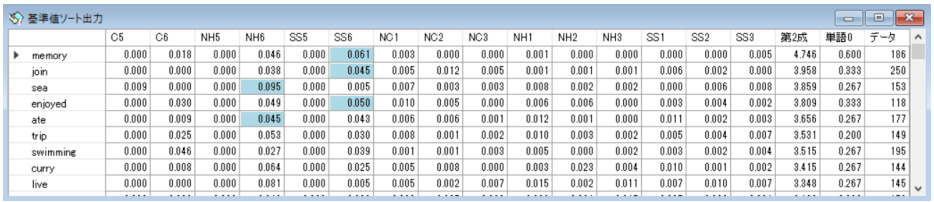
基準値行列出力

基準値行列	C5	C6	NH5	NH6	SS5	SS6	NC1	NC2	NC3	NH1	NH2	NH3	SS1	SS2	SS3
C5	0.155	0.068	0.070	0.058	0.083	0.074	0.064	0.053	0.054	0.064	0.058	0.053	0.062	0.055	0.053
C6	0.068	0.125	0.078	0.081	0.066	0.080	0.066	0.060	0.055	0.071	0.056	0.057	0.065	0.062	0.053
NH5	0.070	0.078	0.168	0.085	0.095	0.069	0.062	0.052	0.052	0.074	0.055	0.051	0.070	0.059	0.051
NH6	0.058	0.081	0.085	0.151	0.060	0.089	0.062	0.057	0.053	0.067	0.058	0.054	0.064	0.058	0.052
SS5	0.083	0.066	0.095	0.060	0.165	0.062	0.060	0.050	0.047	0.059	0.046	0.043	0.061	0.047	0.043
SS6	0.074	0.080	0.069	0.089	0.062	0.146	0.061	0.054	0.049	0.061	0.052	0.050	0.060	0.055	0.048
NC1	0.064	0.066	0.062	0.062	0.060	0.061	0.087	0.063	0.062	0.070	0.066	0.062	0.073	0.065	0.064
NC2	0.053	0.060	0.052	0.057	0.050	0.054	0.063	0.079	0.070	0.064	0.073	0.073	0.060	0.070	0.069
NC3	0.054	0.055	0.052	0.053	0.047	0.049	0.062	0.070	0.081	0.062	0.074	0.077	0.060	0.070	0.075
NH1	0.064	0.071	0.074	0.067	0.059	0.061	0.070	0.064	0.062	0.094	0.067	0.066	0.075	0.068	0.066
NH2	0.058	0.056	0.055	0.058	0.046	0.052	0.066	0.073	0.074	0.067	0.089	0.077	0.064	0.075	0.077
NH3	0.053	0.057	0.051	0.054	0.043	0.050	0.062	0.073	0.077	0.066	0.077	0.094	0.063	0.073	0.077
SS1	0.062	0.065	0.070	0.064	0.061	0.060	0.073	0.060	0.060	0.075	0.064	0.063	0.088	0.063	0.063
SS2	0.055	0.062	0.059	0.058	0.047	0.055	0.065	0.070	0.070	0.068	0.075	0.073	0.063	0.086	0.071
SS3	0.053	0.053	0.051	0.052	0.043	0.048	0.064	0.069	0.075	0.066	0.077	0.077	0.063	0.071	0.087

図 4　300 語での基準値行列

この行列の対角成分には黄色､各行の最も小さな値には緑色の網掛けがしてある｡さらに､この表示にはまだ下があり、そこには教科書の基準値を 2 組掛け合わせた場合の 0 比率が表示されている。この 0 比率が非対角成分の下がり方に影響を与えている。

このデータの場合、第 1 成分の意味は分かったが、第 2 成分以降は単語との関係で意味が決まる。それを見るための機能が「基準値ソート」ボタンである。このボタンの下のテキストボックスに成分の番号を入力し､「基準値ソート」ボタンをクリックすると図 5 の結果が得られる。ここでは第 2 成分についての結果を表示している。



基準値ソート出力

	C5	C6	NH5	NH6	SS5	SS6	NC1	NC2	NC3	NH1	NH2	NH3	SS1	SS2	SS3	第2成	単語0	データ
memory	0.000	0.018	0.000	0.046	0.000	0.061	0.003	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.005	4.746	0.600	186
join	0.000	0.000	0.000	0.038	0.000	0.045	0.005	0.012	0.005	0.001	0.001	0.001	0.006	0.002	0.000	3.958	0.333	250
sea	0.009	0.000	0.000	0.095	0.000	0.005	0.007	0.003	0.003	0.008	0.002	0.002	0.000	0.006	0.008	3.859	0.267	153
enjoyed	0.000	0.030	0.000	0.049	0.000	0.050	0.010	0.005	0.000	0.006	0.006	0.000	0.003	0.004	0.002	3.809	0.333	118
ate	0.000	0.009	0.000	0.045	0.000	0.043	0.006	0.006	0.001	0.012	0.001	0.000	0.011	0.002	0.003	3.656	0.267	177
trip	0.000	0.025	0.000	0.053	0.000	0.030	0.008	0.001	0.002	0.010	0.003	0.002	0.005	0.004	0.007	3.531	0.200	149
swimming	0.000	0.046	0.000	0.027	0.000	0.039	0.001	0.001	0.003	0.005	0.000	0.002	0.003	0.002	0.004	3.515	0.267	195
curry	0.000	0.008	0.000	0.064	0.000	0.025	0.005	0.008	0.000	0.003	0.023	0.004	0.010	0.001	0.002	3.415	0.267	144
live	0.000	0.000	0.000	0.081	0.000	0.005	0.005	0.002	0.007	0.015	0.002	0.011	0.007	0.010	0.007	3.348	0.267	145

図 5　基準値ソート結果

第 2 成分の大きい順に単語が表示され、基準値の値が示されている。上位 5 つの単語については、最も基準値の大きい教科書の位置が青色に網掛けされている。これらの単語と教科書は互いに似た位置にあり、これを用いて利用者は第 2 成分として影響力の大きな単語及びそれに近い教科書を知ることができる。同様に、第 2 成分の小さい（負の）単語についても基準値の値を知ることができる。

2）の場合

ここでは 1 つの教科書の単語数が多く、他も不揃いな場合を考える。語数調整した場合の文書 0 比率と全単語数との関係を表 4 に与える。

表 4　1 重調整法による文書 0 比率と全単語数との相関係数

語数	50	100	300	500	1000	全単語数
Choice-1	0.000	0.070	0.257	0.388	0.582	466
Dening-1	0.000	0.050	0.150	0.212	0.272	3844
Kanda-p1	0.100	0.260	0.517	0.656	0.800	200
Seisoku-1	0.020	0.090	0.280	0.414	0.581	736

Sunshine-1	0.120	0.180	0.420	0.528	0.662	338
Union-1	0.000	0.020	0.157	0.242	0.399	935
相関係数	-0.489	-0.489	-0.637	-0.701	-0.833	

これによると、利用する単語数が多くなると相関は高くなるが、単語数が少ないと相関が低くなり、0 比率を単語数と関連付けることは次第に難しくなる。ただ、0 比率は切り取られた単語の中でどれだけ満遍なく単語を使っているかを表す指標であり、教科書の「標準性」を表す指標のように考えられる。以下には異論があると思われるが、標準的な教科書は比較的やさしいとも考えられ、0 比率は難易度とも関係しているように思われる。ここでは 0 比率を教科書の単語数や標準性を通して難易度と関係する指標と考えて先に進む。

次に、基準値平均について１）の場合に述べたことが成立するか調べてみる。基準値平均については、表 5 の通りである。

表 5　基準値平均とその標準偏差

	50	100	300	500	1000
Choice-1	0.0579	0.0380	0.0199	0.0145	0.0091
Dening-1	0.0502	0.0333	0.0165	0.0127	0.0095
Kanda-p1	0.0555	0.0384	0.0208	0.0138	0.0075
Seisoku-1	0.0569	0.0379	0.0197	0.0146	0.0086
Sunshine-1	0.0524	0.0383	0.0216	0.0164	0.0106
Union-1	0.0514	0.0361	0.0195	0.0146	0.0105
標準偏差	0.0032	0.0020	0.0017	0.0012	0.0012

これによると教科書による標準偏差は基準値平均のほぼ 10%以内に収まっている。また、対角指標については表 6 の関係が得られる。

表 6　対角指標

	50	100	300	500	1000
Choice-1	0.2093	0.2016	0.1885	0.1684	0.1271
Dening-1	0.2172	0.2118	0.1993	0.2004	0.2156
Kanda-p1	0.2645	0.2407	0.1960	0.1436	0.0853
Seisoku-1	0.2204	0.2194	0.2120	0.1890	0.1421
Sunshine-1	0.1924	0.2289	0.2189	0.1972	0.1550
Union-1	0.1842	0.1916	0.1872	0.1808	0.1721

この指標についての全体の平均は 0.1920、標準偏差は 0.0350 である。

次に、主要統計量間の相関行列を求めてみよう。図 6a に 100 語の場合、図 6b に 500 語の場合を与える。

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.908	0.991	-0.906	-0.331	0.103
基準値偏差	0.908	1.000	0.948	-0.718	-0.210	0.168
aii	0.991	0.948	1.000	-0.881	-0.289	0.064
第1成分	-0.906	-0.718	-0.881	1.000	0.054	0.049
第2成分	-0.331	-0.210	-0.289	0.054	1.000	0.003
第3成分	0.103	0.168	0.064	0.049	0.003	1.000

図 6a　主要統計量間の相関行列（100 語）

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.925	0.937	−0.206	0.925	0.258
基準値偏差	0.925	1.000	0.976	−0.071	0.927	0.274
aii	0.937	0.976	1.000	0.066	0.965	0.270
第1成分	−0.206	−0.071	0.066	1.000	0.021	−0.007
第2成分	0.925	0.927	0.965	0.021	1.000	0.041
第3成分	0.258	0.274	0.270	−0.007	0.041	1.000

図 6b　主要統計量間の相関行列（500 語）

100 語では文書 0 比率と第 1 成分とが強い相関を持っているが、500 語ではむしろ第 2 成分の相関が高い。第 3 成分についてはどちらも相関が高くない。そこで、文書 0 比率を第 1 成分と第 2 成分で重回帰分析することを試みる。図 7a は 100 語、図 7b は 500 語の場合である。いずれも重回帰分析の結果と CR 分析による散布図を上下に示している。Dening-1, Kanda-p1, Sunshine-1 の位置を考えるとこれらの結果から、軸が回転している（反転も含む）ことが分かる。

偏回帰係数と検定

文書0比率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値
第1成分	−0.0753	−0.8908	0.0154	−4.8871	3	0.0164
第2成分	−0.0228	−0.2823	0.0147	−1.5485	3	0.2193
切片	0.1175	0.0000	0.0151	7.7761	3	0.0044
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値
	0.949	0.901	0.913	0.834	13.5932	0.0313

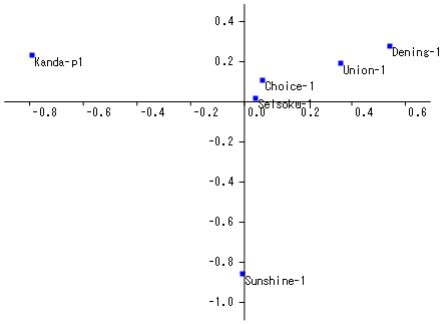


図 7a　重回帰分析と CR 分析の散布図（100 語）

偏回帰係数と検定

文書0比率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値
第1成分	−0.0359	−0.2263	0.0279	−1.2849	3	0.2890
第2成分	0.1442	0.9299	0.0273	5.2804	3	0.0132
切片	0.4163	0.0000	0.0272	15.3180	3	0.0006
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値
	0.952	0.907	0.919	0.845	14.6286	0.0284

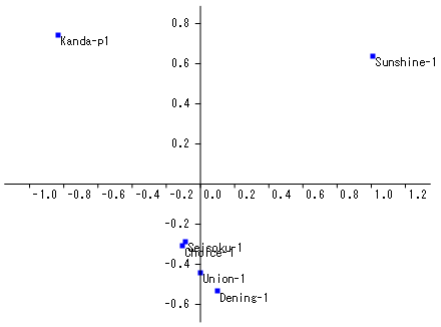


図 7b　重回帰分析と CR 分析の散布図（500 語）

重回帰分析の結果より、第 1 成分と第 2 成分の役割を変えると文書 0 比率をかなりの精度で説明していることが分かる。ではこの回転はなぜ起きるのだろうか。「基準値ソート」ボタンの下のテキストボックスを第「1」成分順にして、「基準値ソート」ボタンをクリックした結果を図 8a（100 語）と図 8b（500 語）に示す。

基準値ソート出力

	Choice-1	Dening-1	Kanda-p1	Seisoku-1	Sunshine-1	Union-1	第1成分	単語0比率	データ順位
had	0.004	0.090	0.000	0.002	0.000	0.048	2.504	0.333	92
was	0.035	0.126	0.000	0.035	0.000	0.038	1.947	0.333	35
when	0.011	0.061	0.000	0.013	0.009	0.052	1.942	0.167	85
as	0.047	0.069	0.000	0.014	0.000	0.057	1.762	0.333	57
out	0.030	0.041	0.000	0.013	0.000	0.054	1.717	0.333	94
of	0.032	0.156	0.010	0.015	0.042	0.071	1.690	0.000	16
so	0.023	0.057	0.000	0.009	0.026	0.043	1.631	0.167	73

図 8a　基準値ソート（100 語）

基準値ソート出力

	Choice-1	Dening-1	Kanda-p1	Seisoku-1	Sunshine-1	Union-1	第1成分	単語0比率	データ順位
pat	0.000	0.000	0.000	0.000	0.117	0.000	3.284	0.833	88
program	0.000	0.000	0.000	0.000	0.113	0.000	3.284	0.833	95
jim	0.000	0.000	0.000	0.000	0.095	0.000	3.284	0.833	129
m	0.000	0.000	0.000	0.000	0.095	0.000	3.284	0.833	129
hi	0.000	0.000	0.000	0.000	0.085	0.000	3.284	0.833	157
oka	0.000	0.000	0.000	0.000	0.080	0.000	3.284	0.833	174
doesn	0.000	0.000	0.000	0.000	0.067	0.000	3.284	0.833	214

図 8b　基準値ソート（500 語）

これを見ると、100 語では標準的な単語が上位を占めているが、500 語では Sunshine-1 で使われている現代的な単語が上位を占めている。一般的な単語は殆どの教科書で使われるので、100 語の場合は「標準性」即ち 0 比率が変動の主流になり、500 語の場合のように特別な単語が特定の教科書で使われている場合は、それらの単語と教科書が変動の主流になる。これが第 1 成分と第 2 成分の交代が起きる理由である。このことから、成分の意味にとって単語の選択数は重要な意味を持っていることが分かる。

３）の場合

ここでは教科書の単語数にほとんど違いがない場合を考える。語数調整した場合の文書 0 比率と全単語数との相関関係を表 7 に与える。

表 7　1 重調整法による文書 0 比率と全単語数との相関係数

語数	50	100	300	500	1000	全単語数
Choice-1	0.020	0.030	0.187	0.332	0.547	466
Drill-1	0.020	0.050	0.200	0.350	0.549	505
J&B-1	0.000	0.080	0.257	0.362	0.506	613
National-1	0.020	0.040	0.200	0.350	0.580	426
Taisho-1	0.000	0.030	0.190	0.316	0.495	633
Tsuda-p1	0.020	0.090	0.260	0.406	0.601	469
相関係数	-0.953	0.049	0.136	-0.352	-0.897	

これによると、利用する単語数が多くなるとやはり相関は高くなるが、そうでない場合、文書 0 比率は単語数にほとんどよらないようである。

次に、基準値平均について１）の場合に述べたことが成立するか調べてみる。結果は表 8 の通りである。

表 8　基準値平均とその標準偏差

	50	100	300	500	1000
Choice-1	0.0551	0.0398	0.0203	0.0146	0.0087
Drill-1	0.0527	0.0357	0.0187	0.0135	0.0087
J&B-1	0.0540	0.0337	0.0173	0.0134	0.0096
National-1	0.0550	0.0397	0.0212	0.0152	0.0089
Taisho-1	0.0514	0.0334	0.0187	0.0141	0.0095
Tsuda-p1	0.0510	0.0353	0.0199	0.0148	0.0093
標準偏差	0.0018	0.0028	0.0014	0.0007	0.0004

教科書による標準偏差は基準値平均の 10%以内に収まっている。また、対角指標については表 9 の関係が得られる。

表 9　対角指標

	50	100	300	500	1000
Choice-1	0.1861	0.2075	0.1853	0.1620	0.1172
Drill-1	0.1997	0.2063	0.1910	0.1658	0.1297
J&B-1	0.2014	0.1854	0.1708	0.1643	0.1501
National-1	0.1851	0.2048	0.1927	0.1660	0.1165
Taisho-1	0.1903	0.1858	0.1823	0.1700	0.1417
Tsuda-p1	0.1826	0.2002	0.2029	0.1775	0.1326

この指標についての全体の平均は 0.1751、標準偏差は 0.0258 である。

以上の結果から、基準値平均についてはほぼ近似が成り立っていると考えることができるが、対角指標については今の段階では何とも言えない。一般に標準化を行わない場合、このようなことはなく、アニメーションで見た結果の安定性も十分ではない。これらの指標と安定性の問題について今後もう少し考察を進める必要があるだろう。

次に、主要統計量間の相関行列を求めてみよう。図 9a に 100 語の場合、図 9b に 500 語の場合を与える。

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.754	0.342	0.694	0.276	0.058
基準値偏差	0.754	1.000	0.002	0.986	-0.024	0.111
aii	0.342	0.002	1.000	-0.149	0.152	0.809
第1成分	0.694	0.986	-0.149	1.000	0.015	-0.008
第2成分	0.276	-0.024	0.152	0.015	1.000	0.024
第3成分	0.058	0.111	0.809	-0.008	0.024	1.000

図 9a　主要統計量間の相関行列（100 語）

主要統計量間相関

	文書0比率	基準値偏差	aii	第1成分	第2成分	第3成分
文書0比率	1.000	0.672	0.813	-0.085	0.214	-0.149
基準値偏差	0.672	1.000	0.784	0.671	0.062	-0.166
aii	0.813	0.784	1.000	0.293	0.555	-0.305
第1成分	-0.085	0.671	0.293	1.000	0.000	0.011
第2成分	0.214	0.062	0.555	0.000	1.000	-0.010
第3成分	-0.149	-0.166	-0.305	0.011	-0.010	1.000

図 9b　主要統計量間の相関行列（500 語）

100 語では文書 0 比率と第 1 成分とがある程度相関を持っているが､500 語ではもはやどの成分とも相関は低い。そこで、文書 0 比率を第 1 成分と第 2 成分で重回帰分析することを試みる。図 10a は 100 語、図 10b は 500 語の場合である。いずれも重回帰分析の結果と CR 分析による散布図を上下に示している。

偏回帰係数と検定

文書0比率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値
第1成分	0.0164	0.6898	0.0092	1.7850	3	0.1723
第2成分	0.0063	0.2653	0.0091	0.6864	3	0.5418
切片	0.0546	0.0000	0.0091	5.9768	3	0.0094
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値
	0.743	0.552	0.503	0.253	1.8483	0.2998

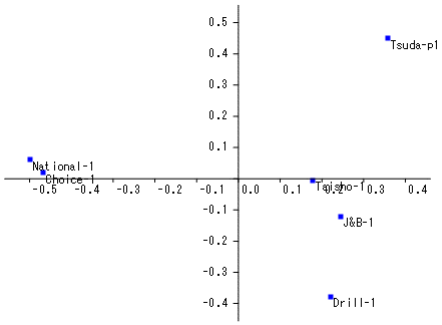


図 10a　重回帰分析と CR 分析の散布図（100 語）

偏回帰係数と検定

文書0比率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値
第1成分	-0.0024	-0.0848	0.0159	-0.1510	3	0.8896
第2成分	0.0060	0.2144	0.0158	0.3816	3	0.7282
切片	0.3528	0.0000	0.0158	22.3695	3	0.0002
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値
	0.231	0.053	0.000	0.000	0.0842	0.9213

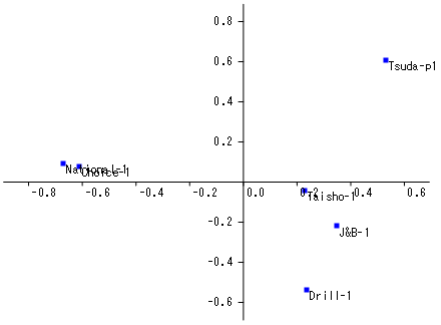


図 10b　重回帰分析と CR 分析の散布図（500 語）

100 語ではある程度の寄与率はあるが、500 語では重回帰式は全く意味がない。以上のように単語数に差がない場合は、文書 0 比率と単語数の相関もないし、成分との関係も得られない。

7.　おわりに

著者らは CR 分析を用いた文書の分析で専用のプログラムを作り、何が成分（軸）の意味を表しているのか、ということを調べてきた。その結果、大きな要素の 1 つは単語数の多さや教科書の標準性に関係する文書 0 比率であった。しかし、この指標も殆ど同じレベルの教科書間では分類に影響を与えない。CR 分析で意味のあることは 0 比率がどの程度分析に影響を与えているのか、また影響を与えているならどの成分が 0 比率を表しているのかを知り、その他の成分の役割を検討することであると思われる。

今回のプログラム作成で未解決な部分は、特に基準値平均が文書によらなかった理由とそれが分析に与える影響である。また、対角指標と呼んだ基準値分散に関係する指標が、文書や切り取った単語数から独立かどうかの見極めも未解決である。さらに、これらは平均的な文章を扱う教科書独自の性質なのか、ある程度一般の文書でも成り立つ性質なのかということも疑問として残っている。今後多くの文書について当たっていけば結論はおのずと見えてくるが、この性質に理論的な説明を付けるのは難しそうである。

参考文献

［1］　福井正康・渡辺清美、「コレスポンデンス分析を用いた英文テキスト分類における語数調整法と単語の選択基準」、福山平成大学経営研究、第 15 号（2019）63-78

［2］　福井正康, 渡辺清美、「テキストコレスポンデンス分析専用プログラムの開発」、日本言語教育 ICT 学会研究紀要、第 7 号、(2020) 49-58

補遺　テキスト CR 分析の理論

教科書ごと単語ごとの出現数のデータを $n_{i\lambda}$ （$1 \le i \le p$，$1 \le \lambda \le m$，$p = m$）とする（調整済みを含む）。ここに p は教科書の数、m は利用する単語の数である。

各文書にパラメータ u_i、各単語にパラメータ v_λ を与え、これを用いて文書と単語の相関係数 ρ を以下のように定義する。

$$\rho = \frac{S_{uv}}{S_u S_v}$$

ここに、

$$S_{uv} = \frac{1}{n}\sum_{i=1}^{p}\sum_{\lambda=1}^{m} n_{i\lambda} u_i v_\lambda, \quad S_u^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i\cdot} u_i^2, \quad S_v^2 = \frac{1}{n}\sum_{\lambda=1}^{m} n_{\cdot\lambda} v_\lambda^2$$

$$n_{i\cdot} = \sum_{\lambda=1}^{m} n_{i\lambda}, \quad n_{\cdot\lambda} = \sum_{i=1}^{p} n_{i\lambda}, \quad n = \sum_{i=1}^{p}\sum_{\lambda=1}^{m} n_{i\lambda}$$

であり、パラメータについては以下を仮定する。

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{p} n_{i\cdot} u_i = 0, \quad \overline{v} = \frac{1}{n}\sum_{\lambda=1}^{m} n_{\cdot\lambda} v_\lambda = 0$$

この相関係数 ρ について、$S_u^2 = 1$，$S_v^2 = 1$ とする制約条件を付けて最大値を求める。そのために Lagrange の未定乗数法を用いる。

$$L = S_{uv} - \alpha\left(S_u^2 - 1\right) - \beta\left(S_v^2 - 1\right)$$

ここに α と β は未定乗数である。この L を u_i と v_λ で微分して、以下の方程式を得る。

$$\sum_{\lambda=1}^{m} n_{i\lambda} v_\lambda - 2\alpha n_{i\cdot} u_i = 0, \quad \sum_{i=1}^{p} n_{i\lambda} u_i - 2\beta n_{\cdot\lambda} v_\lambda = 0$$

左の式に u_i をかけて i について和をとると $\rho = 2\alpha$、右の式に v_λ をかけて λ について和をとると $\rho = 2\beta$ を得る。すなわち、

$$\sum_{\lambda=1}^{m} n_{i\lambda} v_\lambda - \rho n_{i\cdot} u_i = 0, \quad \sum_{i=1}^{p} n_{i\lambda} u_i - \rho n_{\cdot\lambda} v_\lambda = 0$$

次に、右式を v_λ について解いて、

$$v_\lambda = \frac{1}{\rho n_{\cdot\lambda}} \sum_{j=1}^{p} n_{j\lambda} u_j$$

これを左式に代入すると、

$$\sum_{j=1}^{p}\sum_{\lambda=1}^{m} \frac{n_{i\lambda} n_{j\lambda}}{n_{i\cdot} n_{\cdot\lambda}} u_j - \rho^2 u_i = 0$$

さらに、$u_i = \sqrt{n/n_{i\cdot}}\, z_i$ とすると、$\sum_{i=1}^{p} z_i^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i\cdot} u_i^2 = S_u^2 = 1$ となり、以下を得る。

$$\sum_{j=1}^{p} a_{ij}\, z_j - \rho^2 z_i = 0 \qquad \text{(A1)}$$

ここに a_{ij} は以下となる。

$$a_{ij} = \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda}}{\sqrt{n_{i\mathrm{g}} n_{\mathrm{g}\lambda}}} \frac{n_{j\lambda}}{\sqrt{n_{j\mathrm{g}} n_{\mathrm{g}\lambda}}} \right) = \sum_{\lambda=1}^{m} x_{i\lambda} x_{j\lambda} \qquad \text{(A2)}$$

ここに、

$$x_{i\lambda} \equiv \frac{n_{i\lambda}}{\sqrt{n_{i\mathrm{g}} n_{\mathrm{g}\lambda}}} \qquad \text{(A3)}$$

今後 $x_{i\lambda}$ をデータ $n_{i\lambda}$ に対する基準値、a_{ij} が与える行列 **A** を基準値行列と呼ぶ。一般に基準値行列 a_{ij} には以下の関係がある。

$$a_{ij} = \sum_{\lambda=1}^{m} x_{i\lambda} x_{j\lambda} \le \frac{1}{2}\left(\sum_{\lambda=1}^{m} x_{i\lambda}^2 + \sum_{\lambda=1}^{m} x_{j\lambda}^2 \right) = \frac{1}{2}\left(a_{ii} + a_{jj} \right)$$

これらの関係を使うと v_λ は、z_j を用いて以下のようにも書ける。

$$v_\lambda = \frac{1}{\rho n_{\mathrm{g}\lambda}} \sum_{j=1}^{p} n_{j\lambda} u_j = \frac{1}{\rho} \sum_{j=1}^{p} \sqrt{\frac{n_{j\mathrm{g}}}{n_{\mathrm{g}\lambda}}} x_{j\lambda} u_j = \frac{1}{\rho n_{\mathrm{g}\lambda}} \sum_{j=1}^{p} \sqrt{\frac{n}{n_{j\mathrm{g}}}} n_{j\lambda} z_j = \frac{1}{\rho} \sqrt{\frac{n}{n_{\mathrm{g}\lambda}}} \sum_{j=1}^{p} x_{j\lambda} z_j$$

(A1) 式は行列 **A** の固有方程式である。但し、a_{ij} にはその形に起因した以下の制約がある。

$$\sum_{j=1}^{p} a_{ij} \sqrt{n_{j\mathrm{g}}} = \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda}}{\sqrt{n_{i\mathrm{g}} n_{\mathrm{g}\lambda}}} \sum_{j=1}^{p} \frac{n_{j\lambda} \sqrt{n_{j\mathrm{g}}}}{\sqrt{n_{j\mathrm{g}} n_{\mathrm{g}\lambda}}} \right) = \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda}}{\sqrt{n_{i\mathrm{g}} n_{\mathrm{g}\lambda}}} \sum_{j=1}^{p} \frac{n_{j\lambda}}{\sqrt{n_{\mathrm{g}\lambda}}} \right)$$

$$= \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda} \sqrt{n_{\mathrm{g}\lambda}}}{\sqrt{n_{i\mathrm{g}} n_{\mathrm{g}\lambda}}} \right) = \sum_{\lambda=1}^{m}\left(\frac{n_{i\lambda}}{\sqrt{n_{i\mathrm{g}}}} \right) = \sqrt{n_{i\mathrm{g}}}$$

よって、**A** には固有値 1 の自明な固有ベクトル

$${}^t\mathbf{z} = \left(\sqrt{n_{1\mathrm{g}}/n} \quad \sqrt{n_{2\mathrm{g}}/n} \quad \cdots \quad \sqrt{n_{p\mathrm{g}}/n} \right)$$

が存在する。

これは **u** にすると ${}^t\mathbf{u} = \left(1 \quad 1 \quad \cdots \quad 1\right)$ になり、$\overline{u} = (1/n)\sum_{i=1}^{p} n_{i\mathrm{g}} u_i = 1 \neq 0$ であり、平均が 0 の条件を満たさない。また、v_λ についても以下となり、全く特徴を表さない。

$$v_\lambda = \frac{1}{\rho n_{\mathrm{g}\lambda}} \sum_{j=1}^{p} n_{j\lambda} u_j = \frac{1}{n_{\mathrm{g}\lambda}} \sum_{j=1}^{p} n_{j\lambda} = 1$$

そのため、CR 分析ではこの解は省いて表示する。

Multi-purpose Program for Social System Analysis 42
- Text Correspondence Analysis -

Masayasu FUKUI[*1] and Kiyomi WATANABE[*1]

**1 Department of Business Administration, Faculty of Business Administration, Fukuyama Heisei University*

Abstract: The authors of the paper have named a type of correspondence analysis (CR analysis) which analyzes words appeared in a text and their frequencies to examine the similarities among texts as “text CR analysis”. This paper renders a detail explanation of the text CR analysis, which is a part of a statistical analysis software, College Analysis. The program consists of three parts: the standard correspondence analysis part, the part which shows results in a scatter diagram and an animated diagram, and the part that studies what the dimensions produced by CR analysis would mean. The current study particularly focuses on the third part using sample data.

Key Words: College Analysis, correspondence analysis, document analysis