



## コーパス言語学研究における多変量解析手法の比較 ：主成分分析 vs. コレスポネンス分析

著者	水本 篤
雑誌名	統計数理研究所共同研究レポート232 「コーパス言語研究における量的データ処理のための統計手法の概観」
ページ	53-64
発行年	2009
その他のタイトル	Comparison of Multivariate Data Analysis Methods in Corpus Linguistics: Principal Component Analysis vs. Correspondence Analysis
URL	<a href="http://hdl.handle.net/10112/12990">http://hdl.handle.net/10112/12990</a>

## コーパス言語学研究における多変量解析手法の比較 —主成分分析 vs. コレスポンデンス分析—

水本 篤

流通科学大学

E-mail: atsushi@mizumot.com

**あらまし** 本研究では、コーパス言語学において、多変量解析の手法として使用されることの多い、(a) 主成分分析、(b) コレスポンデンス分析、(c) 行列データを転置して行う主成分分析の3つの結果を比較した。その結果、コレスポンデンス分析が変数間の差異を調べるような研究ではふさわしいということが確認されたが、目的に応じて使い分けることが重要であるということも示唆された。

**キーワード** 多変量解析, 主成分分析, コレスポンデンス分析

## Comparison of Multivariate Data Analysis Methods in Corpus Linguistics: Principal Component Analysis vs. Correspondence Analysis

Atsushi MIZUMOTO

University of Marketing and Distribution Sciences (Assistant Professor)

**Abstract** This paper reports on a comparison of multivariate data analysis methods in corpus linguistics: principal component analysis, correspondence analysis, and principal component analysis with transposed matrix. The results show that correspondence analysis is the most suitable method for exploring the differences among variables.

**Keyword** multivariate analyses, principal component analysis, correspondence analysis

### 1. はじめに

コーパス言語学においては、著者推定論やテキスト類型論のための多変量解析手法として、主成分分析とコレスポンデンス(対応)分析の2つを使用することが多い(田畑, 2004b)。主成分分析、コレスポンデンス分析(数量化 III 類と数理的に違いはない分析法)ともに、目的は分析対象としている多数の変数の持つ情報を、できるだけ少ない次元(成分)に圧縮し表現することである。2つの手法とも、結果を2次元(もしくは Nakamura & Sinclair, 1995 のように3次元)でグラフに表し、対象としている変数やサンプルの相対的な位置関係を確認することにより、その特徴を明らかにする。

2つの手法の大きな違いは、対象とするデータの前提条件であり、コレスポンデンス分析は質的データ（名義尺度，順序尺度）を扱い，主成分分析では量的データ（間隔尺度，比率尺度）を扱うとされている。コーパスにおける語やコロケーションの頻度のようなデータは，尺度が等間隔になっているような量的データではなく，質的データであるためコレスポンデンス分析のほうがふさわしいと考えられるが<sup>1</sup>，扱っているデータがコーパスにおける高頻度語のような頻度が適当なレンジにまたがり（頻度が1や2の観測しかないというわけではない），相関係数が計算できるようなデータであれば，記述目的として主成分分析を適応することも可能であるため（出村 他，2004; Tabachnick & Fidell, 2006），主成分分析をコレスポンデンス分析と同じような目的で使うこともある<sup>2</sup>。Baayen (2008) は，“For two-way contingency tables, correspondence analysis provides an attractive alternative. Like principal components analysis, correspondence analysis seeks to provide a low-dimensional map of the data” (p.129). と述べており，金 (2007) は，「対応分析は，データの構造を再現する面では主成分分析より効果が劣るが，パターンを分類する面では主成分分析より良い結果を示すケースが多い」(p. 87) としているが，これら2つの手法によって得られる結果がどのように違うのかについて詳しく説明している文献は少ない。

主成分分析とコレスポンデンス分析の結果の違いとして，主成分分析では，第1主成分に最も説明力の高い成分を抽出し，すべての変数に重みづけをした合成得点（総合指標）を算出するため変数間の差異がわかりにくい。一方，コレスポンデンス分析は，行と列の項目の相関（正準相関）が最大になるように並べ替えて，関連性が強いものやパターンが似ているものが近くになるような値を与える方法であり，行と列の両方に対して標準化をするため，主成分分析での第1主成分にあたる次元（軸）を含まない。つまり，変数間の類似関係のみが強調される（君山，2008）。また，「第2主成分，第3主成分がコレスポンデンス分析の第1次元，第2次元に対応することになる」（君山，2002, pp. 30-31）ともいわれている。これらの違いをまとめると，データ中の変数やケース間の差異を発見することが目的ではなく，できるだけ元の情報を圧縮したものを作り上げることにあるならば主成分分析がすぐれており，差異（類似）を見つけ出すことが目的なのであれば，コレスポンデンス分析が適していると考えられる。

これら2つの方法に加えて，コーパス言語学研究で使われている解析手法として「行列データを転置して行う主成分分析」が存在する。これは，Burrows (1989) 以降，使用されている方法であり，「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」（田畑，2004a）ことで知られている。図1はそのような行列データの転置を示したものである。図1からわかるように，(a)元のデータ行列では，変数の数がケースの数よりも少ないが，(b)データ行列を転置させたものの場合，ケースの数が変数の数よりも多くなっている。このようなデータ配列の場合，相関係数行列（分散共分散行列）が正定値行列ではなくなるので（特異行列），因子分析は実行できないが，主成分分析は実行可能である。このようにして行列を転置した形のデータに対して主成分分析を適応する形の場合，転置しない場合とはかなり違う結果が得られるため，同じ「主成分分析を行った」といっても，違う方法になっているので注意が必要である。コレスポンデンス分析の場合は，行列を入れ替えても結果は同じになるので，このような問題は生じない。

田畑 (2004b) は，Dickens と Smollett の作品における -ly 副詞を対象として，主成分分析とコレスポンデンス分析を行い，2つの結果の比較をした。この研究の主成分分析でも著者間の明確な違いが第2主成分に現われており（第1主成分は合成得点が作られるので分別力がない），コレスポンデンス分析の第1次元に対応していることが明らかになった。また，それらは「行列データを転置して行う主成分分析」の第1主成分に対応していることも確認された。また，生起頻度（粗頻度），相対頻度（100万語あたりなどに換算した場

1 コレスポンデンス分析は「質的データの主成分分析」ともいわれる。

2 大隈 (2006) のように，本来は使うべきではなく，使用する際には注意が必要であるという意見もある。

合の頻度)の両方の場合ともに、コレスポネンス分析のほうが高い弁別性を示しているため、著者推定のような研究ではコレスポネンス分析の方がふさわしいのではないかと  
いう結果を示した。

(a) 元のデータ行列			
	変数 1	変数 2	変数 3
ケース 1	71	62	56
ケース 2	52	58	51
ケース 3	37	48	29
ケース 4	41	66	34
ケース 5	54	41	51

(b) データ行列を転置させたもの					
	ケース 1	ケース 2	ケース 3	ケース 4	ケース 5
変数 1	71	52	37	41	54
変数 2	62	58	48	66	41
変数 3	56	51	29	34	51

図 1 行列データの転置 (サンプル)

水本・野口 (2009) でも、(a)コレスポネンス分析、(b) 通常の主成分分析<sup>3</sup>の第 2 主成分と第 3 主成分、(c) 行列データを転置して行う主成分分析の第 1 主成分と第 2 主成分、の 3 つの結果を比べて、これらは比較的同じような結果を得ることができる方法であるかもしれないということを示した。この研究では、1 つの大規模科学技術英語コーパス (PERC コーパス) における語の頻度と学術領域を対象とした分析を行ったため、同じような比較を別のデータで行った場合、どのような結果が得られるかわからないため、本稿では、上記 3 つの方法で得られる結果が、田畑 (2004b)、水本・野口 (2009) のような先行研究と同じようなものになるのかを調査した。データの分析とその結果の比較、検討は、以下の順で行う。

1. 主成分分析 (特に第 2 主成分と第 3 主成分)
2. コレスポネンス分析
3. 行列データを転置して行う主成分分析

なお、主成分分析は SPSS 14.0 を、コレスポネンス分析は、R version 2.6.2. の MASS ライブラリーの `corresp` 関数を用いた。

<sup>3</sup>主成分分析では、相関行列と分散共分散行列を基に計算する方法の 2 つがあるが、ほとんどすべての先行研究で、相関行列を基にした主成分分析を行っているため、本稿でも相関行列を用いた方法を「主成分分析」と呼ぶ。田畑 (2004b) では分散共分散行列を基に固有値を計算する主成分分析の方法も比べられている。

## 2. 方法

いくつかのコーパスを変数として使用した場合の分析を行うために、表 1 に挙げた 7 つのコーパスを使用した。まず、標準的な英語を含んだものとして、1994 年に完成した総語数 1 億語以上の **British National Corpus** (以下、BNC) から、頻度が 800 以上のものをレマ化してまとめた高頻度語をまとめた **Kilgarriff (1996) の British National Corpus Frequency List** を利用した。その他に、高等学校の英語教科書コーパス (**Senior High**)、センター試験およそ 10 年分をまとめたコーパス (**Center Exam**)、水本 (2004) で作成された **TOEIC** コーパス (**TOEIC**)、洋画のスク립トを集めたコーパス (**Movie**)、英字新聞 (**Washington Post**) のコーパス (**Newspaper**)、世界最大の科学技術英語コーパスである **PERC Corpus** (**PERC**) を対象とした。これらのコーパスを、**JACET8000** (大学英語教育学会基本語改訂委員会, 2003) に収録されている **v8an** というプログラムを使ってレマ化と頻度付与を行った (表 1)。

表 1 から明らかなように、コーパスによって総語数にばらつきがあるので、頻度が付与されている語彙表を作成する際には、すべてのコーパスの総語数を 100 万語に換算した相対化頻度 (**wpm: words per million**) を用いた。そして、完成した語彙表のうち、BNC において機能語を除いた上位 200 語を基準とした。そして、200 語 (行) × 7 つのコーパス (列) のデータ配列を基本の形として分析に用いた。

表 1 分析に使用したコーパスの Type (異語数) と Token (総語数)

コーパス	Type	Token
BNC	N/A	100,000,000 以上
Senior High	12,467	406,032
Center Exam	4,648	90,376
TOEIC	12,825	226,495
Movie	32,213	1,194,163
Newspaper	108,888	4,367,093
PERC	865,266	27,289,174

Note. BNC はもともとリストになっているものを使用したため type は N/A。

表 2 は語の相関係数行列を示したものである。この表の数値を確認すると、TOEIC と PERC が他のコーパスと比べると低い値を示しており、その他のコーパスは相互に高い相関があることがわかる。特に PERC はとても低い相関係数になっていることから、BNC を基にした上位 200 語と PERC コーパスでのそれらの語の使用を比較した場合、科学技術英語である PERC コーパスは語の使い方が大きく違うことがわかる。

表 2 分析に使用したコーパスの相関係数行列

	BNC	Senior High	Center Exam	TOEIC	Movie	Newspaper	PERC
BNC	—						
Senior High	.83	—					
Center Exam	.83	.91	—				
TOEIC	.42	.33	.39	—			
Movie	.74	.67	.75	.28	—		
Newspaper	.74	.65	.55	.37	.42	—	
PERC	.17	.06	.10	.18	-.08	.07	—

### 3. 結果

表 3 は主成分分析の結果である。合成された主成分をどこまで採用するかは、固有値の最小値が 1 という基準が使われることが多いが、今回は、後で第 2 主成分と第 3 主成分を他の手法（コレスポネンス分析，行列データを転置して行う主成分分析）と比べるため，同時に提示している。図 2 は第 1 主成分と第 2 主成分をプロットしたものである。第 1 主成分は元のデータを最も多く吸収しているものであり，表 2 において他のコーパスとの相関が高い（相関和が高い）順番で並んでいることがわかる。

表 3 主成分分析の結果

	第 1 主成分	第 2 主成分	第 3 主成分
BNC	.94	.04	.09
Center Exam	.92	-.07	.14
Senior High	.92	-.11	.13
Movie	.80	-.31	.11
Newspaper	.76	.06	-.16
PERC	.13	.90	.40
TOEIC	.51	.42	-.71
固有値	4.08	1.10	0.75
寄与率	58.30	15.78	10.64
累積寄与率	58.30	74.08	84.72

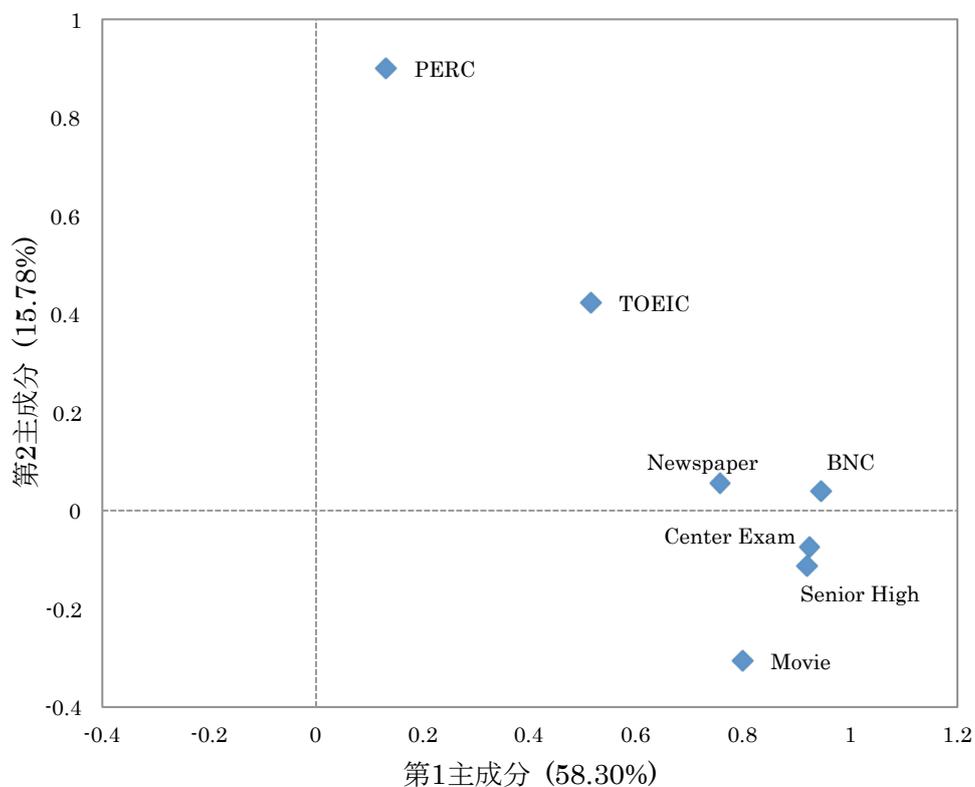


図 2 主成分分析の結果（第 1 主成分と第 2 主成分）

また第2主成分は PERC コーパスの主成分負荷量が一番高く、Movie コーパスが一番低くなっているため、書きことばと話しことばのジャンルの違いを表しているのではないかと考えられる。前掲の相関係数行列をまとめている表2と図3を見比べてみると、この図では、コーパス（変数）間の相関係数を圧縮した形で2次元に表現できていることがわかる。

次に、本研究の目的である、(a) 主成分分析の第2主成分と第3主成分、(b) コレスポネンス分析、(c) 行列データを転置して行う主成分分析、の3つの結果を比較する。図3は主成分分析の第2主成分と第3主成分を散布図にプロットしたものである。前述のように、第2主成分は書きことばと話しことばのジャンルの違いを表していると考えられるが、第3主成分では、PERC コーパスと TOEIC コーパスが成分のプラスとマイナスで分かれている。また、Newspaper コーパスが他のコーパスと比べて、第3主成分では若干マイナスに位置していて、この主成分のマイナスに大きな負荷量を示している語は、company, office などであることから「ビジネス関連の語」が多くなっていて、プラスに大きな負荷量を示している語は、study, result, effect などの語であり、研究や実験に関連していると思われるため、「科学英語で使われやすい語」と考えられる。

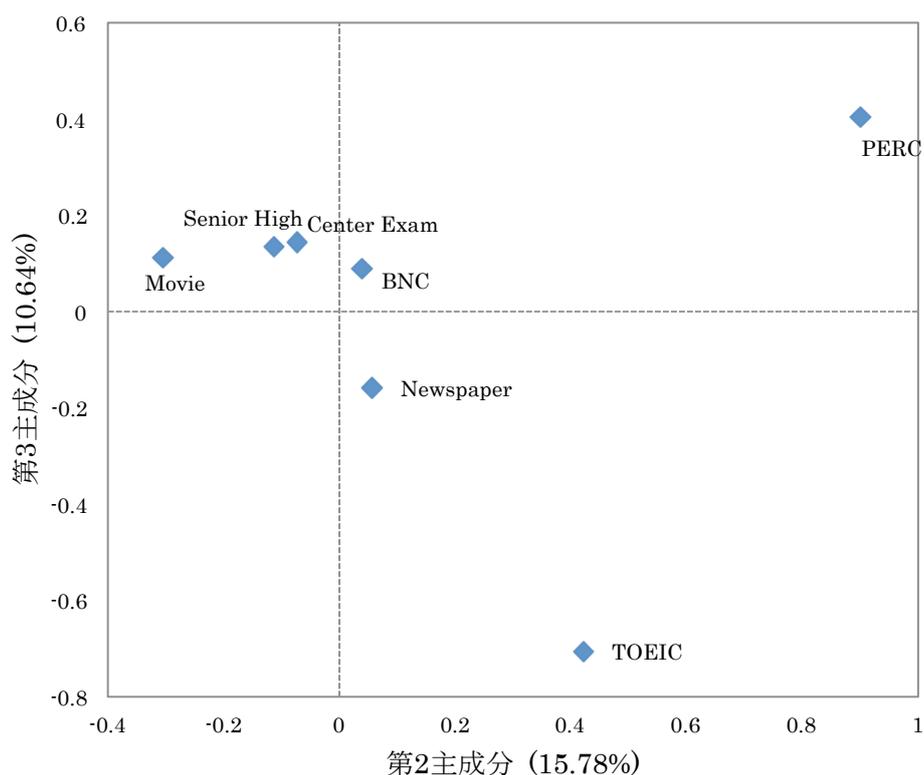


図3 主成分分析の結果（第2主成分と第3主成分）

次にコレスポンデンス分析の結果を表 4, および図 4 に提示する。次元 2 のプラスとマイナスの値を入れ替えば<sup>4</sup>, 図 3 は主成分分析の第 2 主成分と第 3 主成分を散布図にプロットしたものとほぼ同じ結果が, 図 4 のコレスポンデンス分析でも得られることがわかる。

表 4 コレスポンデンス分析の結果

	次元 1	次元 2
Movie	1.31	-0.28
Senior High	0.50	-0.43
Center Exam	0.39	-0.40
BNC	-0.19	-0.09
TOEIC	-0.91	2.03
Newspaper	-0.31	0.64
PERC	-2.73	-2.12
単相関係数 (特異値)	.41	.30
固有値	0.17	0.09
寄与率	44.24	24.00
累積寄与率	44.24	68.24

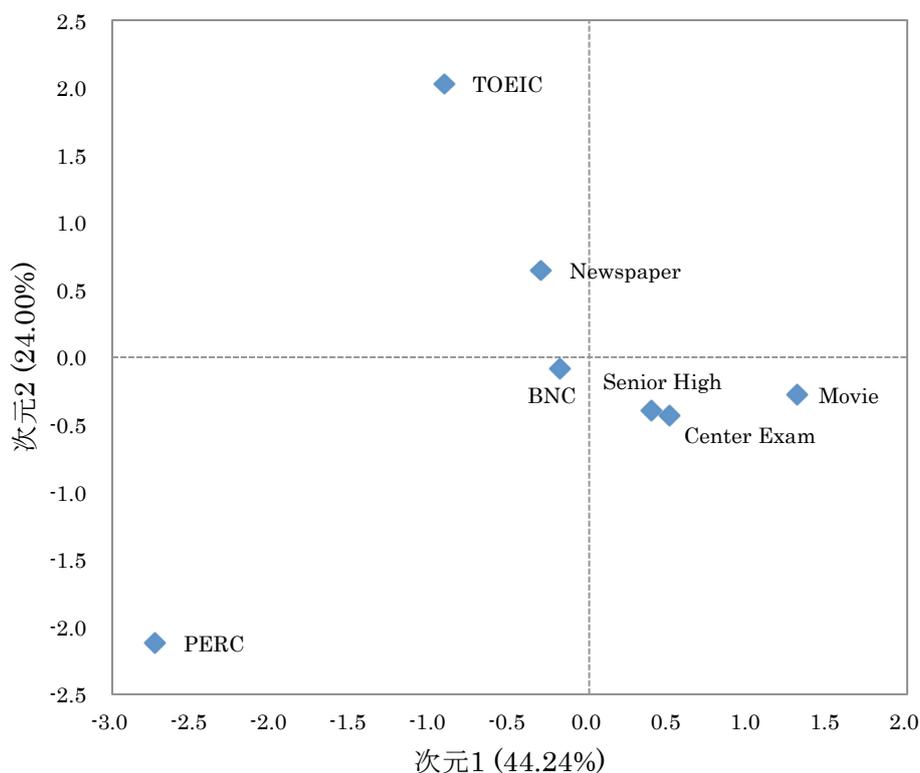


図 4 コレスポンデンス分析の結果

<sup>4</sup> 主成分分析やコレスポンデンス分析では, 負荷量のプラス・マイナスは重要な問題ではない。相対的な位置関係が分析において重視される。

最後に、200 語（行）× 7 つのコーパス（列）のデータ配列を転置して、7 つのコーパス（行）× 200 語（列）にして行う主成分分析の結果を表 5 と図 5 に示す。コレスポンデンス分析の結果（図 4）と比べてみると、それぞれのコーパスは、ほぼ同じ場所に位置している。唯一の違いは、行列データを転置して行う主成分分析では、第 1 主成分に一番高い負荷量を示しているのは Movie コーパスではなく、Senior High コーパスになっていることぐらいである。

表 5 行列データを転置して行う主成分分析の結果（主成分スコアの方を示したもの）

	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分	第 5 主成分	第 6 主成分	第 7 主成分
BNC	-0.17	-0.06	0.64	1.24	1.20	1.32	0
Senior High	1.19	-0.51	0.74	-0.47	-1.41	0.84	0
Center Exam	0.89	-0.28	0.68	-0.66	1.22	-1.37	0
TOEIC	-0.64	1.84	0.00	-1.10	0.01	0.38	0
Movie	0.75	-0.09	-2.12	0.22	0.15	0.11	0
Newspaper	-0.41	0.52	0.25	1.45	-1.02	-1.23	0
PERC	-1.62	-1.42	-0.19	-0.67	-0.16	-0.05	0
固有値	85.43	36.98	31.19	23.52	13.88	9.00	0
寄与率	42.71	18.49	15.59	11.76	6.94	4.50	0
累積寄与率	42.71	61.20	76.80	88.56	95.50	100	100

Note. 固有値が計算されるのは第 6 主成分まで。第 7 主成分以降、200 まで固有値は 0 である。

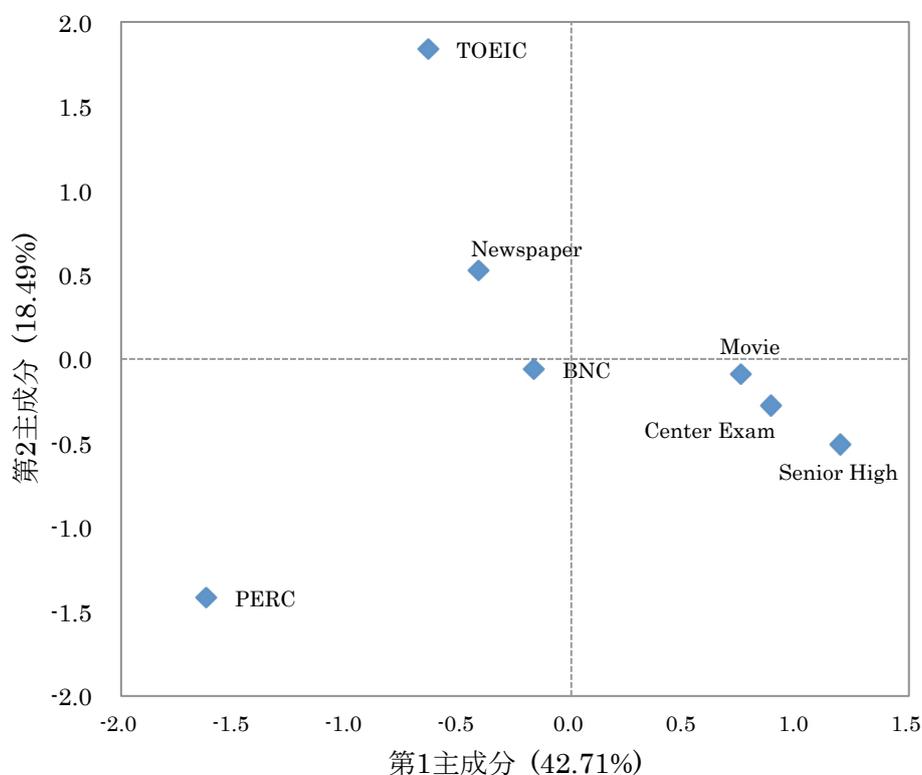


図 5 行列データを転置して行う主成分分析の結果（主成分スコアの方を示したもの）

行列データを転置して行う主成分分析を用いた場合、相関行列の情報は 200 語に対してのものになるため、固有値は 200 が最大となる。ゆえに、第 1 主成分を見てみると、固有値が 85.43 と非常に大きな値になっている。同時に、この手法では、固有値は主成分 7 以降が 0 になる。理論的には主成分は変数の数だけ求めることが可能なので、この場合、200 求めることが可能であるはずだが（語が列の位置にあるため）、行に転置した 7（コーパスの数）から、平均情報 1 を引いた 6 までの次元までの情報しかないためこのような結果になる（廣野・林, 2008）。「行と列の数のうち、少ない方からマイナス 1 だけ固有値が得られる」というのは、コレスポンデンス分析の特徴であるため（高橋, 2005）、行列データを転置して行う主成分分析は、コレスポンデンス分析に近い処理をしているということが推測される<sup>5</sup>。また、変数がサンプル数よりも多くなるので、主成分分析における寄与率は低くなるが、それでもコレスポンデンス分析と同じような結果が得られる方法であることから、変数間（テキスト間）の差異を明らかにしたい研究では使われているのであろうと思われる。

上記までの分析はコーパス（テキスト）間の差異を明らかにすることが目的であったが、それぞれの分析で得られた語に対する負荷量も検討することができる。表 6 は各分析から得られた負荷量上位 10 語をプラスとマイナスで順番に並べたものである。以下が対応している成分（次元・軸）である。

1. 書きことばの特徴を持つ
  - ・主成分分析の第 2 主成分の負荷量がプラス
  - ・コレスポンデンス分析の次元 1 の負荷量がマイナス
  - ・行列を転置した主成分分析の第 1 主成分の負荷量がマイナス
2. 話しことばの特徴を持つ
  - ・主成分分析の第 2 主成分の負荷量がマイナス
  - ・コレスポンデンス分析の次元 1 の負荷量がプラス
  - ・行列を転置した主成分分析の第 1 主成分の負荷量がプラス
3. 科学英語の特徴を持つ
  - ・主成分分析の第 3 主成分の負荷量がプラス
  - ・コレスポンデンス分析の次元 2 の負荷量がプラス
  - ・行列を転置した主成分分析の第 2 主成分の負荷量がマイナス
4. ビジネス英語の特徴を持つ
  - ・主成分分析の第 3 主成分の負荷量がマイナス
  - ・コレスポンデンス分析の次元 2 の負荷量がマイナス
  - ・行列を転置した主成分分析の第 2 主成分の負荷量がプラス

表 6 では、このような 4 つの特徴を 3 つの方法における成分（次元・軸）で比較したときに、同じ語があった場合には太字にしている。数を見ると、全体ではコレスポンデンス分析がもっとも太字の数が多くなっている。つまり、他の 2 つの方法よりも、コレスポンデンス分析は語の特徴をはっきりさせることができる手法であることを裏づけているといえるだろう。

---

<sup>5</sup> 数理的には、「非ゼロの固有値の数が行の数-1になる、というのは相関行列の性質による。相関行列を求める際には変数の平均を引く操作が入るので（z得点の行列に直してから、その転置との積をとって相関行列を求める）、平均を引いた時点で、一次独立なベクトルの数がサンプル数より 1 つ少なくなる」（前田忠彦, personal communication, February 16, 2009）という説明になる。

表 6 それぞれの分析で得られた負荷量上位 10 語

主成分分析			
第 2 主成分 (書きことば⇔話しことば)		第 3 主成分 (科学英語⇔ビジネス英語)	
プラス	マイナス	プラス	マイナス
use	<b>know</b>	<b>study</b>	<b>company</b>
also	go	<b>result</b>	<b>new</b>
<b>result</b>	<b>tell</b>	also	<b>office</b>
time	<b>want</b>	see	year
<b>study</b>	<b>just</b>	<b>effect</b>	<b>question</b>
<b>system</b>	<b>think</b>	then	<b>service</b>
high	<b>thing</b>	use	<b>business</b>
new	<b>come</b>	mean	<b>week</b>
number	say	<b>different</b>	<b>month</b>
<b>effect</b>	people	high	work

コレスポネンス分析			
次元 1 (話しことば⇔書きことば)		次元 2 (科学英語⇔ビジネス英語)	
プラス	マイナス	プラス	マイナス
<b>know</b>	<b>effect</b>	<b>effect</b>	<b>company</b>
yes	<b>result</b>	<b>result</b>	<b>office</b>
<b>tell</b>	<b>level</b>	level	<b>service</b>
<b>just</b>	<b>system</b>	<b>study</b>	<b>business</b>
<b>thing</b>	<b>rate</b>	<b>example</b>	<b>question</b>
right	<b>study</b>	<b>different</b>	<b>market</b>
<b>want</b>	form	particular	<b>month</b>
here	<b>include</b>	<b>case</b>	<b>week</b>
never	<b>require</b>	involve	policy
<b>think</b>	<b>subject</b>	<b>produce</b>	member

行列を転置した主成分分析			
第 1 主成分 (話しことば⇔書きことば)		第 2 主成分 (ビジネス英語⇔科学英語)	
プラス	マイナス	プラス	マイナス
try	<b>system</b>	<b>week</b>	same
hear	set	pay	show
look	<b>rate</b>	<b>month</b>	<b>different</b>
friend	<b>require</b>	<b>business</b>	<b>study</b>
keep	<b>level</b>	<b>office</b>	point
feel	<b>subject</b>	<b>new</b>	<b>example</b>
lot	allow	<b>market</b>	<b>effect</b>
<b>come</b>	<b>case</b>	<b>service</b>	body
great	<b>include</b>	<b>company</b>	<b>case</b>
put	development	<b>question</b>	<b>produce</b>

Note. 同じ色をつけているものが比較対象となる。太字は比較箇所と同じ語が 1 度でも出ているものを示す。

本研究での結果、および先行研究から得られた結果をまとめると、次の表 7 のようになる。(a) 主成分分析の第 2 主成分と第 3 主成分、(b) コレスポネンス分析、(c) 行列データを転置して行う主成分分析、の 3 つの手法はほとんど同じ結果（布置図による解釈）になるがコレスポネンス分析が変数間（もしくはサンプル間）の類似関係や差異をもっとも表すことができる。行列を転置した主成分分析もコレスポネンス分析と同じような結果になるが、目的が類型化などであるなら、行列のデータを入れ替えたとしても結果が変わらないコレスポネンス分析のほうがふさわしいといえるだろう。

表 7 3 つの手法の違いのまとめ

	主成分・次元（軸）			注意すべき点
主成分分析	1	2	3	第 2, 第 3 主成分では寄与率が低くなる
コレスポネンス分析	-	1	2	外れ値に影響される
行列を転置した主成分分析	-	1	2	寄与率がコレスポネンス分析に比べて低くなることが多い
成分や軸に示される特徴⇒	合成 得点	ジャンルの 違い	その他	

#### 4. おわりに

本研究では、コーパス言語学においてよく使用される多変量解析である、(a) 主成分分析 (b) コレスポネンス分析、(c) 行列データを転置して行う主成分分析の 3 つを、いくつかのコーパスから得られた語彙頻度表に対して行った。結果は、先行研究である田畑 (2004b) と同じく、コレスポネンス分析がもっとも類似関係や差異をはっきりと弁別できるというものであった。「行列データを転置して行う主成分分析」が、コレスポネンス分析よりも有用である場合があるとすれば、どのようなデータに対して使うべきなのか、などという点は今回の研究では明らかにされていないため、今後も調査を続けていくべきである。

また、「主成分分析かコレスポネンス分析のどちらを使うべきか？」という疑問に対しては、目的が差異の発見ではなく、データを圧縮して関係性を見る場合には主成分分析、類似関係や差異を見つけることが目的であるならコレスポネンス分析というように、一般的には分類できる。しかし、これらの手法は、多変量のデータに対して探索的に使用し、分割表のみではわからないような視点を提供することができるので、データに合わせて、目的を理解した上で使い分けていくことが必要であるだろう。

#### 謝 辞

本稿の執筆にあたり、Professional English Research Consortium (PERC)、武庫川女子大学の野口ジュディー先生、および、小学館に PERC コーパスのデータを提供していただきました。また、大阪大学大学院言語文化研究科の田畑智司先生に、データ分析方法についてたいへん貴重なアドバイスをいただきました。ここに記して感謝いたします。

## 文 献

- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Burrows, J. F. (1989). "A Vision' as a Revision?" *Eighteenth-Century Studies*, 22, 551-565.
- 大学英語教育学会基本語改訂委員会 (2003). 『JACET8000』 東京: JACET.
- 廣野元久・林 俊克 (2008). 『JMPによる多変量データ活用術 [2訂版]』 東京: 海文堂.
- 出村慎一・西嶋尚彦・長澤吉則・佐藤進 (編). (2004). 『健康・スポーツ科学のためのSPSSによる多変量解析入門』 東京: 杏林書院.
- Kilgarriff, A. (1996). *BNC database and word frequency lists*. Retrieved October 8, 2006, from <http://www.kilgarriff.co.uk/bnc-readme.html>
- 君山由良 (2002). 『コレスポネンス分析と因子分析によるイメージの測定法』 東京: データ分析研究所.
- 君山由良 (2008). 『データ分析入門 2 多変量解析法・MDSの応用』 東京: データ分析研究所.
- 金 明哲 (2007). 『Rによるデータサイエンス』 東京: 森北出版株式会社.
- 水本 篤 (2004). 「JACET8000とTOEICの相関性: TOEIC ミニコーパスを使つての数量的検討」 『JACET8000活用事例集』 60-61. 東京: 大学英語教育学会.
- 水本 篤・野口ジュディー (2009). 「多変量解析を用いたPERCコーパスの領域分類」 『統計数理研究所共同レポート』
- Nakamra, J., & Sinclair, J. (1995). The world of woman in the Bank of English: Internal criteria for the classification of corpora. *Literacy and Linguistic Computing*, 10, 99-110.
- 大隅 昇 (2006). 「よくある質問とそれへの回答」 『「テキスト型データのマイニングとその応用」 WordMiner 活用セミナー資料』 Retrieved January 25, 2008, from <http://wordminer.comquest.co.jp/wmtips/analysis.html>
- Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th international ed.). Boston, MA: Pearson/Allyn & Bacon.
- 田畑智司 (2004a). 『コーパス言語学のための多変量解析入門』 英語コーパス学会第24回大会ワークショップ配布資料. Retrieved January 22, 2006, from <http://www.lang.osaka-u.ac.jp/~tabata/JAECS2004/JAECS2004hand.pdf>
- 田畑 智司 (2004b). 「-ly 副詞の生起頻度解析による文体識別—コレスポネンス分析と主成分分析による比較研究—」 田畑 智司 (編) 『電子化言語資料分析研究』 (pp. 97-114). 大阪大学言語文化部・大阪大学大学院言語文化研究科.
- 高橋 信 (2005). 『Excelで学ぶコレスポネンス分析』 東京: オーム社.