統計数理研究所共同研究リポート 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』(2009) pp. 85–106

多変量解析を用いた PERC コーパスの領域分類

水本　篤
流通科学大学
E-mail:　atsushi@mizumot.com

野口　ジュディー
武庫川女子大学
E-mail:　khb04356@nifty.ne.jp

あらまし　本研究は, 世界でも最大の科学技術英語コーパスである PERC コーパスの領域(サブ・コーパス）が，どのように分類できるか，多変量解析を用いて検討することを目的とした。PERC コーパスの 22 領域における高頻度語を用い, 機能語を除く上位 200 語を分析対象として，(a) クラスター分析, (b) 因子分析, (c) 主成分分析, (d) コレスポンデンス分析の 4 つを行った。その結果，それぞれの方法で目的に合わせて領域分類が行えることが確認され，PERC コーパスは ESP（EAP）において利用価値が高い言語資料であるということが確認された。
キーワード　多変量解析，PERC コーパス，ESP，EAP，語彙研究

Using Multivariate Data Analyses to Categorize Subcorpora of the PERC Corpus

Atsushi MIZUMOTO
University of Marketing and Distribution Sciences (Assistant Professor)

Judy NOGUCHI
Mukogawa Women's University (Professor)

Abstract　The purpose of current study is to investigate how subcorpora of the PERC Corpus can be categorized into registers by using a multivariate data analysis approach. The most frequent words, excluding function words, were used for the following analyses: cluster analysis, factor analysis, principal component analysis, and correspondence analysis. The results show that the PERC Corpus can be categorized into reasonable groups and that it is a very valuable source of ESP (EAP) discourse.
Keyword　multivariate analyses, PERC Corpus, ESP, EAP, vocabulary research

1. はじめに

1.1. ESP と EAP におけるコーパス言語学の利用

特定の目的のための英語（ESP: English for Specific Purposes）は，ある特定の職業（やそれに関連した目的）において必要な英語を研究対象とした分野として，外国語教育学や応用言語学における研究に大きな影響を及ぼしてきたと考えられている（Dudley-Evans, 2001)。ESP の最大の特徴は，指導と教材がニーズ分析（needs analysis）の結果に基づいているということであり，必要な活動に適した，文法，語彙，レジスター[1]，スキル，ディスコース[2] などを中心に学習することになる（Dudley-Evans & St John, 1998)。このような ESP のアプローチは海外においては 30 年間以上研究されており，日本国内の高等教育機関においても，近年，ESP 教育に対する関心は非常に高まっている（田地野, 2004; 深山, 2000 など)。

ESP はさらに，学術目的の英語（EAP: English for Academic Purposes）と職業目的の英語（EOP: English for Occupational Purposes）の 2 つに大きく分類することができる (Dudley-Evans, 2001)。EAP，そして ESP 全般において，目標としているところは，ある特定（職業）領域の言語を使用している専門家集団（ディスコース・コミュニティー）の一員となることであり（田地野, 2004)，そのために，専門家集団が使用している言語のテクスト分析が重要な位置を占める。ゆえに，*Journal of English for Academic Purposes* や *English for Specific Purposes* のようなジャーナルの論文においては，コーパスを用いた研究は必要不可欠なものになっている。例えば，ある分野における研究者が，学術論文においてどのようなどういう書き方をしているかということを理解することによって，研究者になることを目指す者が，そのモデルを模倣することができる。そのようなテクスト分析では，コーパスを利用することでより具体的な言語材料を抽出することができるため，コーパス言語学は ESP 研究において貴重な情報を提供している。

Coxhead（2000）は，商業・法学・自然科学・人文科学の 28 分野から成る 3,600,000 語のアカデミック・コーパスを作成し，すべての分野に万遍なく現れるワードファミリー（happy, unhappy, happiness, unhappiness, happily を 1 語とカウントする方法）で 570 語の Academic Word List (AWL) を作成した。AWL は学術論文のようなアカデミックな英文においては，少なくとも全体の 8.5% はカバーできるため（Ming-Tzu & Nation, 2004)，その有用性は Chen and Ge (2007) や Vongpumivitch, Huang, and Chang (2009) でも実証されており，EAP においては必須となる語彙を集めたリストである。日本では，石川 (2005) が 238,000 語の米国司法文献コーパスから，特徴語を抽出し，256 語の司法英語 ESP 語彙表を作成するなど，ESP（EAP）語彙リストの作成はコーパス言語学を利用することによって可能になった研究分野であるといえる。

---

[1] レジスター（register）とは，言語を使用する状況に応じて，語彙や文法，発音などを変えることによって生じる言語の様々な種類のことを指す。一番わかりやすいレジスターは，書きことばと話しことばの違いである。レジスターは「言語使用域」と呼ばれることもある。

[2] 1 つの文や個人の 1 回の発言を超えた，もっと大きなレベルでの言語の構成組織。やりとりの過程の中で意味交渉が行われる。ディスコースの研究は談話分析（discourse analysis）と呼ばれる。

単語レベル以外でも多くの研究がおこなわれている。Gledhill (2000) は約 50 万語の医学論文（がん研究）のイントロダクションにおける高頻度コロケーションを分析し，いかにそのような分野では限定された，予測しやすい成句的表現を使っているかを明らかにした。また, 最近では, チャンクやクラスターと同じ意味で使われている lexical bundles (“as can be seen” や “or something like that” のようなかたまり）が，書きことば，話しことばの両方においてディスコースを示す重要な役割を担っていると考えられており，コーパスを使った研究が進んでいる (Biber, 2006; Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Hyland, 2008a; 2008b)[3]。

コーパスを利用した ESP（EAP）の研究においては，分析の基となるコーパスがどのようなものであるかが，結果の一般化には必要となるが，ESP（EAP）においては，1 億語のイギリス英語を収録している British National Corpus のように大規模コーパスを利用することは難しい。また, Coxhead (2000) で使用された Science のサブコーパスが Biology, Chemistry, Computer science, Geography, Geology, Mathematics, Physics の 7 領域のみであったことからも，特定の分野を網羅的にカバーするコーパスが必要であったといえる。

1.2. PERC コーパス

上述の問題点を克服する，科学技術・理工学分野に特化したコーパスが，Professional English Consortium (PERC) が作成した PERC コーパスである（2008 年 6 月より公開)。PERC コーパスは，著作権使用許可を得た約 1,700 万語の学術雑誌論文（1995 年～2002 年）から成るコーパスで，ライフサイエンスを含む，技術，工学，理化学分野 22 領域が含まれている[4]。これだけ特定の分野にターゲットを絞った大規模なコーパスは，公開されている中では世界でも最大の科学技術英語コーパスであり，EAP，ESP における研究では利用価値の高いコーパスである。そこで本研究では，多変量解析を用いて PERC コーパスの領域がどのように分類可能であるかを検討することを目的とした。

2. 方法

2.1. PERC コーパス語彙頻度表のレマ化とデータ行列の作成

まず, 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。このプログラムは JACET8000 に基づいて，分析対象コーパスの中で使われている語をレマ化[5]し，レベル 1（1000 位までの 1000 語）からレベル 8（8000 位）までのランク付けとカバー率を算出するものである。JACET8000 のリストに含まれていない語は “over8” となり, その他, 省略形 (cont. forms)

[3] 引用している Biber や Hyland の研究では, four-word lexical bundles を分析対象としており, その理由を Hyland (2008b) は“they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles” (p. 8) と述べている。

[4] PERC Corpus Online (https://www.corpora.jp/~perc04/) の説明に基づく。その他の詳細は PERC のホームページ (http://www.perc21.org/) を参照。

[5] レマ化とは屈折形を原形にまとめること。

や固有名詞（proper nouns），そして数字などの語としてはカウントされないもの（non-words)がこのプログラムによって自動的に分類された。表 1 は，PERC コーパス 22 領域における Type（異語数）と Token（総語数）をまとめたものである。

表 1 からわかるように，領域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（wpm: words per million）相対化頻度を用いた。そして，完成した語彙表のうち機能語を除いた上位 200 語を分析の対象とした。

表 1　PERC コーパス 22 領域における Type（異語数）と Token（総語数）

領域	日本語	Type	Token
Agriculture	農業	37,871	929,563
Biology	生物学	84,125	2,498,374
Chemistry	化学	90,123	2,615,871
Civil Engineering	土木工学	23,789	622,595
Computer Science	コンピュータサイエンス	69,598	2,881,692
Construction Building Technology	建築・建造	4,049	44,648
Earth Science	地球科学	64,242	2,266,101
Electrical Electronic Engineering	電気・電子工学	33,993	1,090,339
Engineering	工学	75,012	2,911,616
Environmental Sciences	環境科学	5,461	45,019
Fisheries	漁業	17,869	375,560
Food Science	食品科学	25,264	451,603
Forestry	林業	7,434	113,599
General Science	科学一般	27,178	499,525
Materials Science	材料学	28,956	653,307
Mathematics	数学	13,999	367,729
Medicine	医学	151,141	6,215,952
Metallurgy Metallurgical Engineering	金属学・金属工学	9,439	147,978
Nuclear Science Technology	原子力工学	12,691	207,672
Oceanography	海洋学	18,064	457,792
Physics	物理学	55,661	1,724,376
Telecommunications	通信工学	9,307	168,263
Total		865,266	27,289,174

Note. 現在公開されている完成版 PERC コーパスでは，総語数は約 1,700 万語であるが完成前のデータを使用したため，総語数が約 2,700 万語と多くなっている。

	A	B	C	D	E	F	G	H	I	J	K
1	RANK	WORD	Agriculture	Biology	Chemistry	Civil_Engineering	Computer_Science	[illegible]	Earth_Science	[illegible]	Engineering
2	1	figure	1946.08	3830.09	2629.72	3207.54	2274.36	2844.47	3550.59	2623.95	2954.03
3	2	eq	93.59	248.96	1073.83	2390.00	1990.50	851.10	616.48	3160.48	3153.92
4	3	et	55.94	5431.53	1383.10	2569.89	728.39	447.95	3043.11	192.60	1764.31
5	4	use	2176.29	1687.50	2222.97	2622.89	3101.65	2127.75	1664.09	3106.37	2773.72
6	5	al	37.65	5411.92	1320.40	2542.58	731.86	425.55	3011.78	186.18	1621.09
7	6	model	1537.28	550.76	831.85	4518.19	2534.62	1097.47	2118.18	1593.08	2840.69
8	7	high	2428.02	1172.76	1490.13	1299.40	784.96	3785.16	1879.00	1487.61	1397.51
9	8	result	1635.18	1559.01	1686.25	2171.56	1564.71	2665.29	1736.02	1515.13	2110.86
10	9	also	1651.31	1779.16	1634.64	1513.02	1595.94	2262.14	1726.75	1833.37	1566.14
11	10	used	1594.30	1366.49	1594.12	2137.83	1695.88	2306.93	1292.97	2000.30	2258.20
12	11	value	1488.87	555.16	1433.56	2592.38	1576.50	2486.11	1722.34	1387.64	2069.98
13	12	time	1270.49	1108.32	1205.33	2508.85	2599.51	2015.77	1580.69	1847.13	1903.75
14	13	data	1601.83	1187.57	1007.31	2221.35	2517.27	627.13	2027.27	1679.29	1740.96
15	14	study	1954.68	1679.09	1608.64	1482.50	885.94	1455.83	1447.86	509.93	1225.78
16	15	low	1752.44	945.82	1188.90	1286.55	599.30	2194.95	1795.15	1017.11	1071.91
17	16	sample	1863.24	972.23	1298.23	2301.66	444.88	2351.73	1781.47	631.00	1037.91
18	17	system	1417.87	627.21	1298.61	1625.45	2765.74	358.36	1055.56	2939.45	2284.30
19	18	effect	1625.49	1410.52	1446.55	1214.27	629.84	1164.67	846.39	732.80	1215.48
20	19	only	1170.44	1156.35	1094.47	1196.60	1575.12	918.29	1062.62	1385.81	1267.34
21	20	number	1027.36	968.23	706.46	1101.84	2056.08	537.54	749.75	1389.48	1284.51
22	21	temperature	1276.94	443.49	1213.74	613.56	129.09	1343.85	1990.64	785.99	1329.50
23	22	rate	1513.61	589.18	1006.93	925.16	645.11	1209.46	1068.80	845.61	1226.47
24	23	cell	720.77	5412.72	2112.87	342.12	618.39	22.40	372.45	462.24	844.89

図 1　分析に使用した 200×22 の行列の一部

図 1 は分析に使用した，200 語（行），22 領域（列）をまとめたものである。上位 200 語という分析基準は，今回の研究で用いる多変量解析で安定した結果を得るために設定したものである（因子分析の場合，標本数が推定するパラメータ数の 5〜10 倍であるほうが好ましいとされている。市川，2008)。しかし，コーパスから抽出された語の頻度データは高頻度のものであれば，相関係数がかなり高くなると考えられるため，石川（2007）でも実証されているように，20（語）×7（変数）でも因子分析の結果は，より多くの語を含んだ場合と比べても，ほとんど変わらなかった。また，小林（2007）や 石川（2007）では，コレスポンデンス分析の結果はサンプル数を少なくしていっても，カテゴリーのポジショニングにあまり影響は出ないということが報告されている。ゆえに，本研究においても，200 語ではなく，上位高頻度語から一部を取り出しても，結果はほぼ再現されると考えられる[6]。

表 2 は，分析に使用した上位 200 語の PERC コーパス 22 領域の相関係数行列である。この表で相関が高いもの同士が何であるかを確認するだけでも，ある程度の傾向を確かめることができるが，22 領域すべてを一度に見渡すには複雑すぎる。そのような理由からも，今回の研究で用いているような多変量解析によって，情報を圧縮し整理する必要があることがわかるだろう。

6　実際に 200 語から 50 語まで減らして分析してみると，主成分分析，コレスポンデンス分析での領域の位置関係はほとんど変わらなかったが，クラスター分析や因子分析では，分類される領域が若干異なる結果となった。しかし，全体的にはほぼ同じ結果が得られる。

表 2　PERC コーパス 22 領域の相関係数行列（上位 200 語）

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1. Agriculture	—																					
2. Biology	.21	—																				
3. Chemistry	.35	.59	—																			
4. Civil Engineering	.44	.35	.54	—																		
5. Computer Science	.21	.19	.38	.63	—																	
6. Construction Building Technology	.30	.10	.47	.40	.11	—																
7. Earth Science	.37	.57	.66	.75	.43	.43	—															
8. Electrical Electronic Engineering	.19	.10	.41	.54	.84	.16	.41	—														
9. Engineering	.30	.37	.63	.82	.78	.33	.70	.77	—													
10. Environmental Sciences	.48	.38	.15	.22	.06	.04	.40	.02	.16	—												
11. Fisheries	.49	.53	.43	.61	.28	.31	.70	.21	.46	.54	—											
12. Food Science	.25	.78	.64	.40	.13	.24	.58	.06	.36	.33	.55	—										
13. Forestry	.58	.46	.30	.53	.08	.23	.59	.04	.36	.60	.65	.49	—									
14. General Science	.36	.36	.36	.35	.44	.12	.44	.37	.42	.47	.43	.28	.24	—								
15. Materials Science	.28	.26	.70	.52	.39	.54	.61	.51	.68	.04	.27	.26	.20	.23	—							
16. Mathematics	-.09	-.01	.13	.28	.36	.01	.05	.46	.46	-.07	.00	-.02	-.03	.06	.27	—						
17. Medicine	.27	.90	.63	.39	.25	.14	.55	.15	.40	.34	.55	.79	.46	.40	.28	-.02	—					
18. Metallurgy Metallurgical Engineering	.18	.22	.56	.37	.24	.43	.59	.35	.48	.08	.23	.20	.22	.17	.75	.09	.20	—				
19. Nuclear Science Technology	.26	.21	.60	.48	.36	.33	.62	.53	.57	.12	.36	.24	.22	.31	.59	.14	.23	.50	—			
20. Oceanography	.27	.37	.41	.67	.31	.33	.77	.31	.53	.26	.57	.31	.49	.32	.43	.01	.34	.40	.43	—		
21. Physics	.13	.25	.62	.65	.61	.26	.56	.74	.81	.07	.28	.25	.18	.30	.68	.65	.26	.54	.64	.42	—	
22. Telecommunications	.15	.11	.24	.39	.77	.03	.22	.71	.55	.03	.17	.01	.02	.27	.27	.34	.18	.15	.24	.14	.46	—

2.2. 使用した多変量解析の手法

今回の研究の目的は PERC コーパスの領域がどのように分類可能であるかを検討することであるので，表 3 にまとめた多変量解析のうち，「複数の変数間の関連性を検討する，圧縮・整理する」ことが目的である，(a) クラスター分析，(b) 因子分析，(c) 主成分分析，(d) コレスポンデンス分析[7] の 4 つを用いた。

表 3　多変量解析の目的と手法，および尺度水準（小塩, 2004 による分類）

目的	多変量解析の手法	尺度水準	
		従属変数 （基準変数，目的変数）	独立変数 （説明変数）
1 つの変数を 複数の変数から 予測・説明・ 判別する	重回帰分析	量的データ	量的データ
	数量化 I 類		質的データ
	判別分析	質的データ	量的データ
	数量化 II 類		質的データ
複数の変数間の 関連性を検討する 圧縮・整理する	クラスター分析	量的データ	
	因子分析		
	主成分分析		
	数量化III類	質的データ	
	コレスポンデンス (対応)分析		

推測統計学的な考察を行う際には，変数の正規性や多変量正規性など，さまざまな前提が分析するデータに備わっているかを確認しなければならない（Tabachnick & Fidell, 2006)。しかし，今回の分析では，扱っているデータがコーパスにおける高頻度語，かつ頻度が適当なレンジにまたがるようなデータであり（頻度が 1 や 2 の観測しかないというようわけではない)，相関係数が計算できるため量的データとみなすことに問題はないと考えられる[8]。このように，データの記述が目的であるときには，頻度データも量的データとして扱うことが多い。以下ではそれぞれの手法を概観する。

2.2.1. クラスター分析

クラスター分析は，似ているものを分類する手法である。分類する対象はサンプル（図 1 では行として並ぶ個々の 200 語）であることが多いが，変数（図 1 では横の 22 領域）であることもある。また類似度を表す尺度として，サンプルのクラスター化が目的のときには距離を，変数のクラスター化が目的のときには相関係数を用いる（出村 他, 2004)。クラスター分析では，(a) サンプル間（もしくは変数間）の距離を用いて，近いもの（類似したもの）を近くに分類し，1 つのクラスターを作成する，(b) できたクラスターを他のクラスターと結合させていく，(c) デンドログラム（ツリー，樹形図）を確認し，どのようなまとま

[7] 数量化III類とコレスポンデンス（対応）分析の計算方法はほぼ 100%同じである (高橋, 2005)。

[8] ただし，因子分析の最尤法（maximum likelihood estimation）のような方法では多変量正規性の前提が満たされている必要がある (Leech, Barret, & Morgan, 2005)。

りになっているかを判断する，という流れで実施する。クラスター分析では，距離の定義方法（ユークリッド距離，標準化ユークリッド距離，マハラノビスの距離，相関係数），そしてクラスターの結合法[9]（最近隣法，最遠隣法，グループ間平均連結法，重心法，メディアン法，ウォード法）がいくつかあるため，組み合わせによって結果が大きく異なる場合がある。また，デンドログラムを用いてどのようなまとまりになっているかを確認する場合も，どこで区切るかによって，1 つのクラスターに属する数が変化することも十分あり得るため，村上 (1994) は，「どの高さで切断するのが妥当であるかが問題となるが，理論的に決める方法はない」(p. 49) と述べている。つまり，研究者が自分で決めるしか方法がないといえるだろう。これらの理由により，出村 他 (2004) は，クラスター分析について，「用いる方法により，異なる結果が得られるので，1 つの方法の結果のみから結論づけるのではなく，クラスター分析以外の多変量解析法を駆使して総合的な判断を下すことが重要である」(p.92) としている。クラスター分析は，量的データのみならず，質的データにも適応できるため (古谷野, 1988)，データがどのように分類できるかを，とりあえず探索的に調べたいときには，はじめに利用してみるのにふさわしい方法であるといえよう。コーパス言語学でクラスター分析を利用した研究は，村上 (1994, 2004) で，誘拐犯の声明文の筆者特定にクラスター分析を適応している例などが紹介されているが，海外のジャーナルでの報告例はあまり見られない。

2.2.2. 因子分析

（探索的）因子分析は，観測された変数に影響を与えている潜在的な因子を探る手法であり，今回の研究では，22 領域の裏側にはどのような要因（因子）が隠れていて，その要因が何であるのかを明らかにするのが目的である。因子分析は直接観測できない潜在因子を探るのが目的であるため，「PERC コーパスの領域がどのように分類可能であるか」という，今回の目的においても，コーパスで観測される頻度情報が，何らかの目には見えない「構成概念」によって引き起こされているということは十分考えうることであり，内的な構造を理解する目的で分析を行った。因子分析の具体的な手順としては，まず，(a) 変数間の相関係数を計算し，(b) 因子の抽出を行い，そして，(c) 因子を解釈しやすいように軸の回転を行う。(d) 最後に得られた因子を解釈する，という流れになる (前田, 2004)。因子分析の手順の中でも，因子の抽出（主因子法，一般化した最小二乗法，最尤法など）や軸の回転（直交回転，斜交回転など）は利用する方法によって異なった結果が得られるため注意が必要である。

コーパス言語学の分野で因子分析を利用した研究としては，Biber (1988) が報道，フィクション，手紙などの 23 の言語使用域（レジスター）と文法を主とする言語項目 67 点を因子分析し，話しことばと書きことばの特徴を明らかにした。また，Biber, Conrad, Reppen, Byrd, and Helt (2002) でも同様のデザインで，大学生の話しことばと書きことばの差異を調査している[10]。安本・本多 (1981) は，現代作家 100 人の作品 100 編における，名詞の長さ，漢字，句読点などの 15 の特徴を変数として用い，因子分析を行い，作家の分類ができることを明らかにした[11]。Nakamura (1995) が批判しているように，因子分析は因子の抽出方法，軸の回転方法などを（データに合わせて研究者が）恣意的に決定するため，誤った解釈をしてしまう可能性もあるが，正しく使い分けることができれば，データの情報をそのまま圧縮している，主成分分析やコレスポンデンス分析で得られる結果とは違った，潜在的な因子を探ることができるため，とても有効な手法になるといえるだろう。

---

[9] これらの結合法のうち，重心法，メディアン法，ウォード法は，ユークリッド距離以外では用いることができない (出村 他, 2004, p. 92)。

[10] Biber et al. (2003) では因子分析が multidimensional (MD) analysis と呼ばれているが，具体的にどのような違いがあるのかは明記されていない。また，因子の抽出法，軸の回転法などについても述べられていない。

[11] 変数間に強い相関がある場合，多重共線性が存在するため，重回帰分析の場合は好ましくないが，本研究で使用されている多変量解析手法ではそのような心配はしなくてもよい。

2.2.3. 主成分分析

主成分分析は，多数の変数の持つ情報を少数個の成分に圧縮することが一番の目的の手法である。手順としては，(a) 変数間の相関行列を基に，どの程度元のデータの情報を保持しているのかを示す固有値（eigenvalue）と，各主成分における係数である固有ベクトル（eigenvector）を計算する，(b) 固有値から寄与率や累積寄与率を求める，(c) 結果の解釈，という単純なプロセスによって行われる。結果の解釈の際には，固有ベクトルに固有値の平方根をかけた主成分負荷量を用いて変数間の視覚化を行うことが多い（田畑, 2004）。

主成分分析は，SPSS のような統計パッケージでは因子分析の中に含まれるため，同じものであると誤解されることが多いが，因子分析とは違い，軸の回転は行わないのが普通である（川本, 2004）。ゆえに，因子分析と主成分分析は似て異なるものである。図 2 に主成分分析と因子分析の違いを図示した。主成分分析は，すべての変数を重みづけした合計点を計算しているイメージである。一方，因子分析は似通った変数どうしをなるべく統合し，異なる変数群どうしは分ける解析法（対馬, 2008）である。このように主成分分析と因子分析は異なった分析目的により実行される。実際に，目的，そして論理がまったく別のものであるが，この 2 つの手法はたいへんよく似ており，神宮・土田（2008）はその違いを以下のようにまとめている。

> この違いは 2 つの点に集約できる。1 つは計算手法という点で，これはとてもよく似ている。統計ソフトによっては，因子分析の共通性の推定手法の選択肢の中に主成分分析（主成分解や非反復主因子法とよばれることもある）が含まれている場合もある。因子分析では，誤差を取り除き分析するために共通性を推定する。しかし，主成分分析では，誤差はない，あるいは無視して計算するので共通性は 1 であり，推定を行わない。また，回転も行わない。実は，これ以外は，因子分析と主成分分析は同じだといえる。(p. 28)

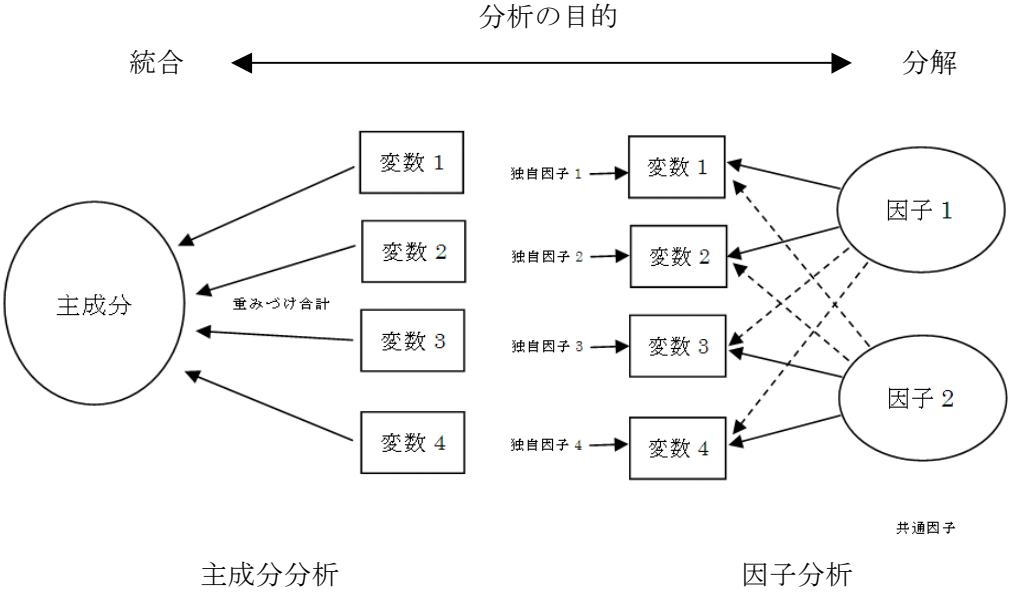

図 2　主成分分析と因子分析の違い

主成分分析では，元のデータ情報の損失を少なくした，最も説明力の高い第 1 主成分を抽出するように分析が行われ，図 2 のようにして求められた第 1 主成分の次に取り出される，第 2 主成分以下の成分は既に得られた成分の全てと直交するように求められるため，得られた第 1 主成分と第 2 主成分は無相関（$r = .0$）となるようになっている。第 1 主成分は元データを最も多く吸収しているものになるので，相関が高いデータの場合には総合得点のようなものを求めていることになる。そのため，変数として使用するデータが相反するものであったり，相関がないものであれば，主成分の計算はできるが，総合得点のようなものは求められない（小椋, 2006）。ゆえに，ある程度の相関が期待される変数を使用すべきである。また，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，分散共分散行列を基にした主成分分析は特殊な場合のみ使われるため，多くが相関行列を基にした主成分分析を行っている。相関行列を使った主成分分析は，すべての変数の分散が 1 になるので，固有値の合計は変数の数と一致する。

主成分分析を使ったコーパス言語学における研究は，Burrows (1989) や Tabata (1995), 田畑 (2005) などをはじめとして，著者推定研究では主要な方法論となっている（田畑, 2005）。主成分分析は同一のデータであれば誰が分析しても同じ結果が出てくると言われているが（朝野, 2007），田畑 (2004) で報告されているとおり，Burrows (1989) 以降，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が採用されているため，行列の配置によって違う結果が得られることに注意しなければならない。

2.2.4. コレスポンデンス分析（対応分析）

コレスポンデンス（対応）分析は，質的データの分析に用い，クロス集計表において，行と列の項目の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る（金, 2007）。より少ない次元に多数の変数をまとめる手法であることから，コレスポンデンス分析は，量的データを対象としている主成分分析と考えられることが多い。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) と述べている。

コレスポンデンス分析は，(a) 行と列の相関係数が最大となるように，固有値（特異値）を求める。(b) 最大固有値は 1 になるので，第 2 固有値と対応する次元を第 1 次元とし，第 2 次元以降も求める（クロス集計表の行列の少ない方マイナス 1 の次元まで求められる），(c) 寄与率と累積寄与率を求める，(d) 固有ベクトルに基づいて（通常 2 次元までの）カテゴリー・スコアとサンプル・スコアを算出し，第 1 次元と第 2 次元にスコアを布置して結果を解釈する，という流れで行う。コレスポンデンス分析は行と列（変数とサンプル）を入れ替えても結果は変わらない「対称性」を持っているので，主成分分析よりも結果が明解である。また，主成分分析では第 1 主成分に総合的指標が抽出されるが，コレスポンデンス分析ではそのようなことはない（廣野・林, 2008）。コレスポンデンス分析は，評価対象の差異が明確になる（強調される）ので，コーパスで変数間やサンプル間の相対的な近さや遠さを見たい時に適している手法であるといえる。しかし，行や列のうちどちらかの合計に外れ値（極端に小さい値）がある場合には，差異が強調されすぎるので注意が必要である。コーパス言語学では，Nakamura and Sinclair (1995), Tabata (2002), 小林 (2007) など数多くの研究でコレスポンデンス分析が用いられている

本研究の分析においては，因子分析，クラスター分析，主成分分析は SPSS 14.0 を，コレスポンデンス分析は, R version 2.6.2. の MASS ライブラリーの corresp 関数を用いた。

3. 結果と考察

3.1. クラスター分析の結果

PERC コーパスにおける 22 領域のそれぞれが，どれだけ近い（遠い）かを調べるために，まずクラスター分析を行った。変数のクラスター化が目的のときには相関係数を用いる（出村 他，2004）ため，最遠隣法，ピアソン相関によるクラスター分析を実行して得られたデンドログラムが図 3 である。実際には，距離では相関係数だけではなく，ユークリッド距離も含み，すべてのクラスターの結合法の組み合わせを試した上で，一番わかりやすい結果になる組み合わせを選んだ[12]。

2.2.1. で説明したとおり，デンドログラムをどこで区切るかについては理論的な確認方法はない。磯田（2004）は，「傾向の似ていないもの同士がつなげられる場合，結合距離が遠くなります。したがって，結合距離が大きく跳ね上がる，つまり，横の線が長くなるところを探すことが方策のひとつです」(p.118) としている。今回の例では，それぞれの領域のまとまりを考えながら，結合距離が 20 辺りで図 3 にあるように，縦線を入れてみて，関係が近いものが何であるのかを検討した。1 と示したクラスターは生物系の領域，2 は工学系が多く，3 は情報工学系の領域が含まれていることがわかる。

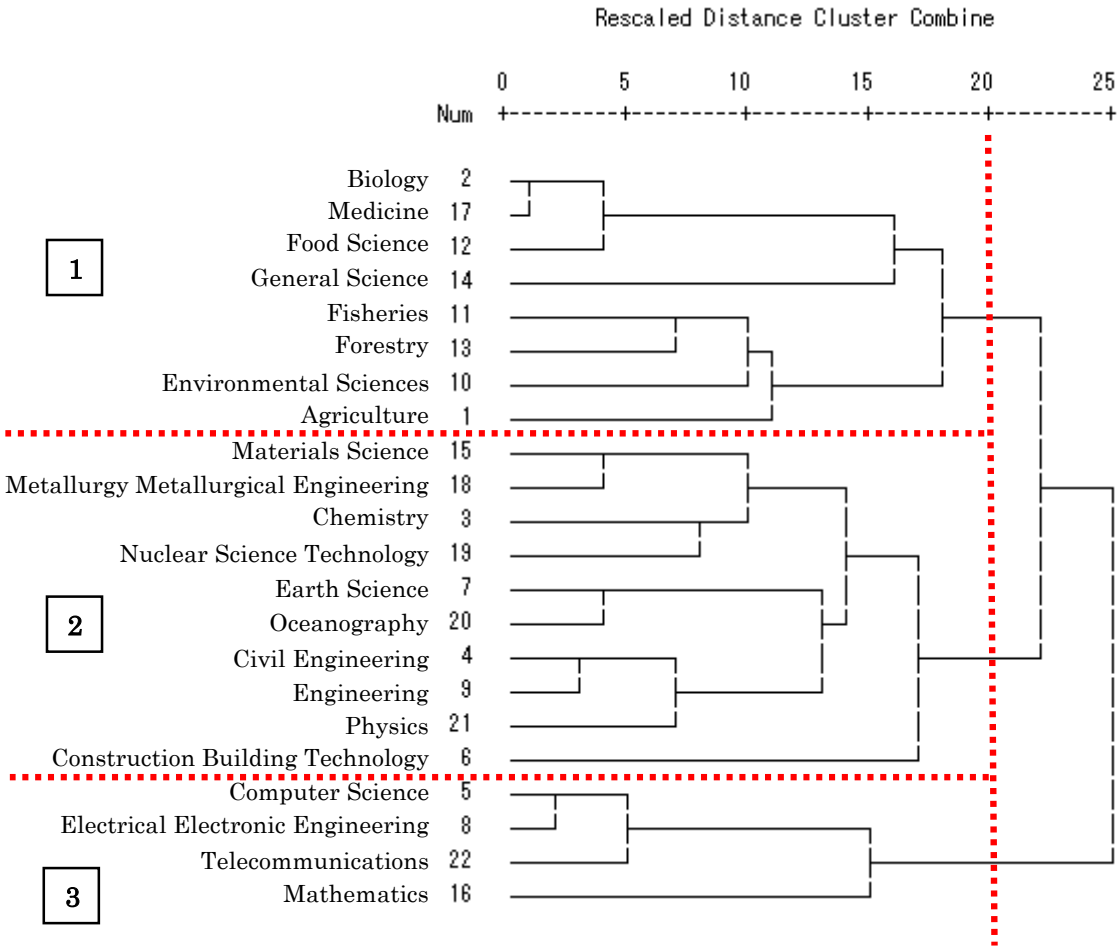

図 3　クラスター分析の結果（最遠隣法，ピアソン相関）

[12] さらに主成分分析，コレスポンデンス分析の結果との比較も行いながら最終的な距離，クラスター結合法を選んだ。

3.2. 因子分析の結果

PERC コーパスの 22 領域に対して，重みなし最小二乗法（プロマックス回転[13]）による因子分析を行った。因子の抽出の基準は最小の固有値を 1 とし（固有値は, 9.09, 3.37, 1.86, 1.54 と変化），表 4 のような 4 因子構造が得られた。回転前の 4 因子で 22 領域のすべての分散を説明する割合は 72.05%であった。SPSS では因子分析を行うのが妥当かどうかを判断する指標として，カイザー・マイヤー・オルキンの標本妥当性（KMO 測度: Kaiser-Meyer-Olkin Measure of Sampling Adequacy）と バートレットの球面性検定 (Bartlett test of sphericity）があり，KMO 測度は 0 ~ 1 の値をとり，0.90≦優良, 0.90≦良好，0.70≦中等度，0.60≦やや不良，0.50≦不良，0.50＜不可,という基準があるが（対馬, 2008），KMO 測度は 0.87 で，バートレットの球面性検定は $p < .001$ であり，因子分析を本研究のデータに適応させることが妥当であることが確認された。

表 4　探索的因子分析の結果（重みなし最小二乗法，プロマックス回転）

	因子 1	因子 2	因子 3	因子 4
第 1 因子:情報工学系分野				
Computer Science	**1.01**	-.22	.08	.01
Electrical Electronic Engineering	**.93**	.06	-.05	-.11
Telecommunications	**.90**	-.29	.00	.01
Engineering	**.65**	.29	.12	.06
Mathematics	**.58**	.03	-.25	-.03
Physics	**.57**	**.49**	-.18	.04
第 2 因子:化学系分野				
Materials Science	.02	**.97**	-.17	-.03
Metallurgy Metallurgical Engineering	-.14	**.90**	-.09	-.06
Construction Building Technology	-.25	**.72**	.13	-.14
Chemistry	.02	**.63**	-.17	**.50**
Nuclear Science Technology	.13	**.62**	.05	-.06
Earth Science	-.02	**.53**	**.44**	.14
第 3 因子:自然科学系分野				
Forestry	-.20	.04	**.84**	.03
Environmental Sciences	-.06	-.25	**.76**	.06
Fisheries	.00	.02	**.74**	.14
Agriculture	-.03	.06	**.70**	-.15
Oceanography	-.01	**.40**	**.46**	-.06
Civil Engineering	**.37**	.28	**.46**	-.07
General Science	.33	-.15	**.38**	.15
第 4 因子:生物系分野				
Medicine	.05	-.14	.01	**.99**
Biology	-.01	-.11	.01	**.98**
Food Science	-.15	.07	.06	**.82**

因子相関行列	因子 1	因子 2	因子 3	因子 4
第 1 因子:情報工学系分野	－			
第 2 因子:化学系分野	.61	－		
第 3 因子:自然科学系分野	.35	.50	－	
第 4 因子:生物系分野	.31	.48	.59	－

Note. 因子負荷量の絶対値が 0.35 以上のものを太字にした。

[13] 直交回転であるバリマックス回転でも試行し，同じ結果が得られた。

4 つの因子は，それぞれに含まれている領域から，第 1 因子を情報工学系分野，第 2 因子を化学系分野，第 3 因子を自然科学系分野，第 4 因子を生物系分野と名づけた。Physic のような領域は，因子負荷量が 2 つの因子にまたがって高くなっており，情報工学系分野（第 1 因子）と化学系分野（第 2 因子）の 2 つに影響を受けていると考えられる。尺度作成のような目的ではこのような 2 つの因子に高い因子負荷量を示している変数（項目）は削除して，再度因子分析を行うのが通例であるが，今回は PERC コーパスで観測された語の頻度に影響を与えている潜在的な因子を探るという，記述的な目的で因子分析を行ったため削除する必要はないと判断できる。因子間の相関係数も中程度（r = .31 ~ .61）となっており，PERC コーパスで使用されている語の使用に影響を与えている潜在因子は，ある程度の相関があると考えられる。

因子分析で得られた結果は，3.1.のクラスター分析の結果と比べてみると，同じような領域分類になっているところもあれば，まったく違う箇所もある。これは，クラスター分析は単純に領域間が近いか遠いかを示しているのに対し，因子分析では領域の近さ（遠さ）だけが問題になっているのではなく，因子分析の際に軸を回転させることにより，構成概念としての因子の解釈をしやすくしているからであると考えられる。例えば，図 3 のクラスター分析の結果では，Mathematics は他の領域と遠い位置にあるが，因子分析では，第 1 因子の情報工学系分野に 0.58 の因子負荷量を示しており，この因子負荷量をもとに 2 次元のプロットを作図したとしても，他の領域との近さや遠さは可視化できない（これは，後述の主成分分析，コレスポンデンス分析と因子分析を比べても同じことが言える）。つまり，因子分析で得られる結果は，現実データをそのまま縮約したものではなく，データの裏側にある目には見えない潜在的な要因を表しているのである。ゆえに，「PERC コーパスで観測される上位 200 語の 22 領域における使用は，情報工学系分野，化学系分野，自然科学系分野，生物系分野という 4 つの隠れた要因によって引き起こされており，これらの領域における学術論文の著者は，このような分野の違いを意識しているため，結果として語の頻度にも違いが出る」ということができる。このような特性を活用し，因子に特徴的な語を調べることもできる。表 5 はそのような各因子で因子スコアの高い語上位 20 語を並べた表である。200 語程度の分析でも，このように分野によって語の使用に違いがみられることがわかる。

表 5　各因子で因子スコアの高い語上位 20 語

順位	第 1 因子： 情報工学系分野	第 2 因子： 化学系分野	第 3 因子： 自然科学系分野	第 4 因子： 生物系分野
1	eq	figure	et	et
2	use	temperature	al	al
3	system	use	figure	cell
4	figure	eq	model	figure
5	model	surface	sample	study
6	time	result	data	protein
7	used	high	area	use
8	data	model	high	also
9	result	value	time	effect
10	number	used	use	result
11	also	also	low	high
12	value	low	rate	level
13	function	sample	study	used
14	only	phase	value	activity
15	case	material	species	data
16	set	structure	water	time
17	process	study	result	sample
18	method	et	also	control
19	different	system	used	low
20	state	solution	temperature	shown

3.3. 主成分分析の結果

図 4 は主成分分析で得られた第 1 主成分と第 2 主成分をプロットしたものである。図中の 1～3 の分類を 3.1.のクラスター分析の結果と比較するために入れた。第 1 主成分は，元データを最も多く吸収しているものになるので，相関が高いデータの場合には総合得点のようなものであると考えられる。つまり，第 1 主成分負荷量が高い Earth Science や Engineering のような領域は，他の領域との相関の高いものであり，上位 200 語では，度の領域ともある程度高い相関係数を示していることが予想できる[14]。逆に，Mathematics や，Environmental Sciences は他の領域との相関係数は低いため，特異な語の使い方をする領域なのであると判断できる（上位 200 語でも頻度 0 の語が存在する）。

クラスター分析の結果と比べると，領域の近さや遠さが近い結果で図示されている。第 1 主成分は前述のように，総合得点のようなものであると考えられるので，第 2 主成分以降に分野として特徴的な語が抽出されている可能性がある。領域の分類については，破線で囲っているクラスター分析の結果と同じ様に，生物系，工学系，情報工学系の領域が近くに位置していることがわかる。

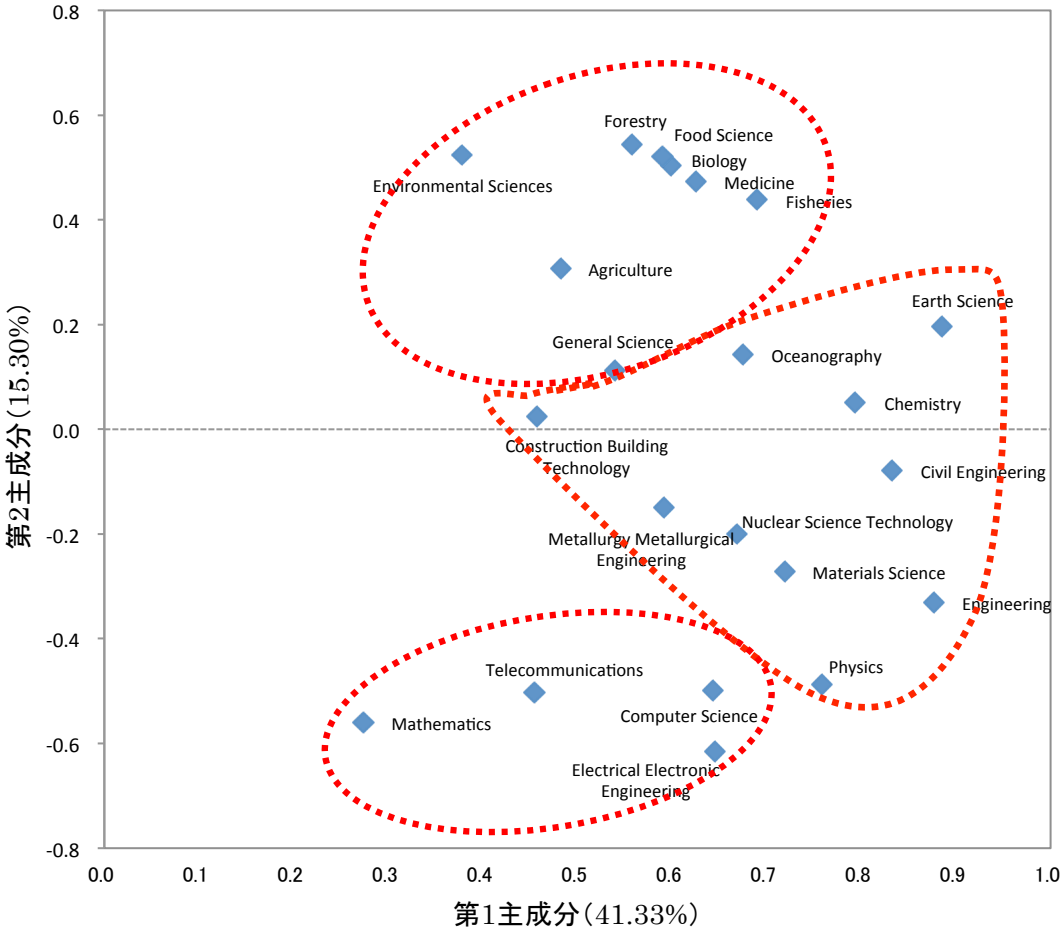

図 4　主成分分析の結果

[14] 実際，表 2 の相関行列の横列の合計（相関和）の順位と，第 1 主成分の主成分負荷量が高い領域は一致する。相関係数を基にした主成分分析の第 1 主成分では必ずそのようになる（菅, 2007)。

図 4 で第 2 主成分が何を示しているのかを解釈すると，プラスの主成分負荷量が生物系分野を示し，マイナスになるにつれて情報工学に関連した領域なっていくと考えられる。第 2 主成分でプラス（生物系）の主成分スコアを示している上位 20 語は，al, et, species, plant, study, area, site, protein, cell, forest, water, day, concentration, treatment, soil, effect, acid, sample, growth, activity，マイナス（情報工学系）の主成分スコアが高くなっている 20 語は，eq, system, network, use, case, section, set, signal, function, then, state, problem, order, point, given, energy, design, application, model, parameter のような語で，この解釈が妥当であることを裏付けている[15]。因子分析と比べると，違った語が抽出されているのが興味深い。これは，主成分分析と因子分析の分析結果が違っていても，朝野 (2007) が述べているように，真実は「どちらも正しい」ため[16]，目的に応じてそれぞれを使うべきであるということを意味している。

3.4. コレスポンデンス分析（対応分析）の結果

図 5 はコレスポンデンス分析の結果をプロットしたものである。主成分分析の結果と比べると，Mathematics や Environmental Sciences が，他の領域と遠い位置にあることが強調された結果となっている（破線での領域分けはクラスター分析の結果に基づく）。

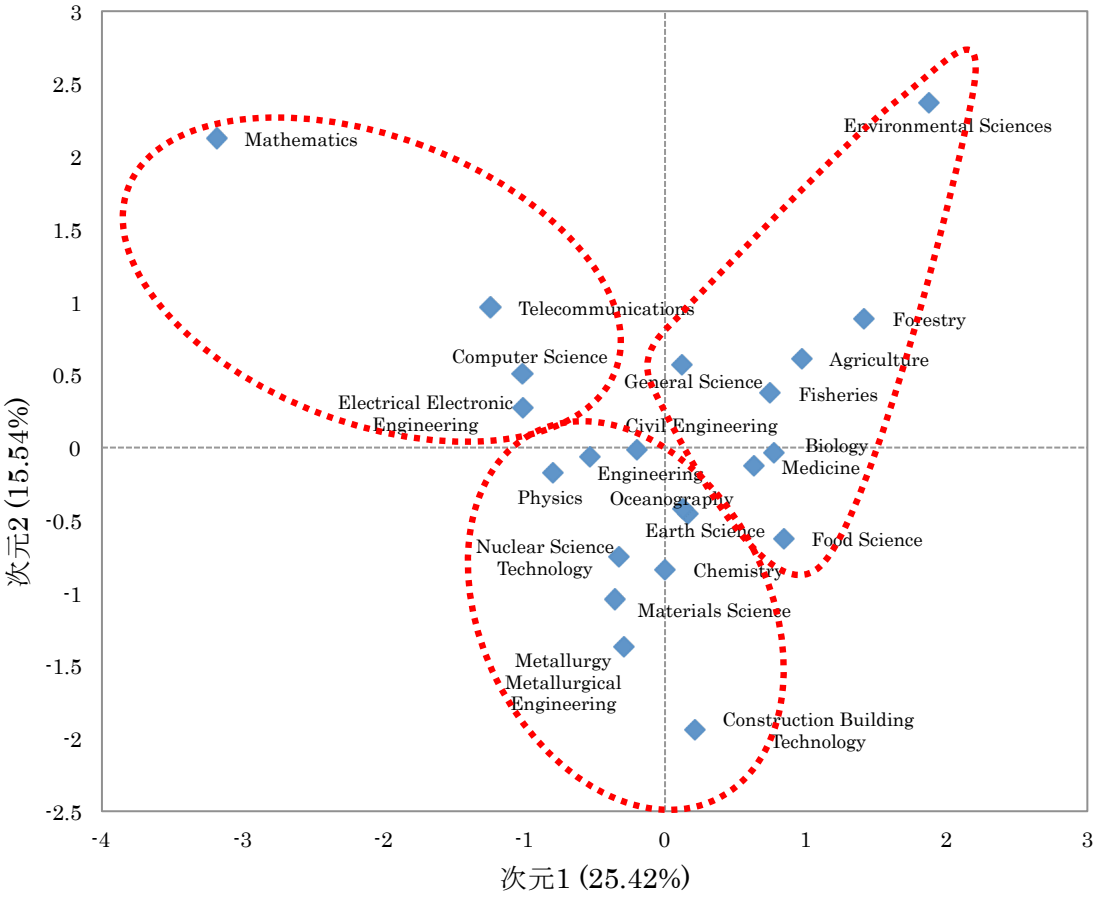

図 5　コレスポンデンス分析の結果（カテゴリー・スコア）

15 主成分分析を行い，主成分得点を算出し，どの語が近いかをクラスター分析で調べる方法も一般的に行われる。

16 すべての分散を説明する割合は 72.05%であった。

領域の分類については，クラスター分析，主成分分析と同様に，生物系，工学系，情報工学系の領域が近くに位置している。図 6 は分析に使用した 200 語（サンプル）をプロットしたものである。この図からもわかるとおり，forest や eq など，ある特定の領域（Mathematics や Environmental Sciences）にのみ非常に高い頻度で出現する語（外れ値）がある場合，コレスポンデンス分析では，このように差異が強調されすぎるという面もある。このような結果が得られたときには，クロス集計表（語彙頻度表）や相関係数行列に戻り，データがどのようなものになっているかを再確認する必要があるだろう。今回のケースでは，forest の場合，Environmental Sciences と Forestry にのみ高頻度で集中的に出現し，Construction Building Technology, Food Science, Materials Science, Mathematics, Metallurgy Metallurgical Engineering, Nuclear Science Technology, Oceanography, Telecommunications の 8 領域では頻度が 0 であったために，このような結果になったのだろうと考えられる。また，コレスポンデンス分析では，次元（成分）は頻度の割合のパターンを強調するので，カテゴリーによって頻度の割合に変化がないものは中心に集まるといわれている（廣野・林，2008）。ゆえに Engineering や Civil Engineering など，主成分分析の第 1 主成分で高い値を示していた領域が，図 5 では中心に集まっている様子がわかる。

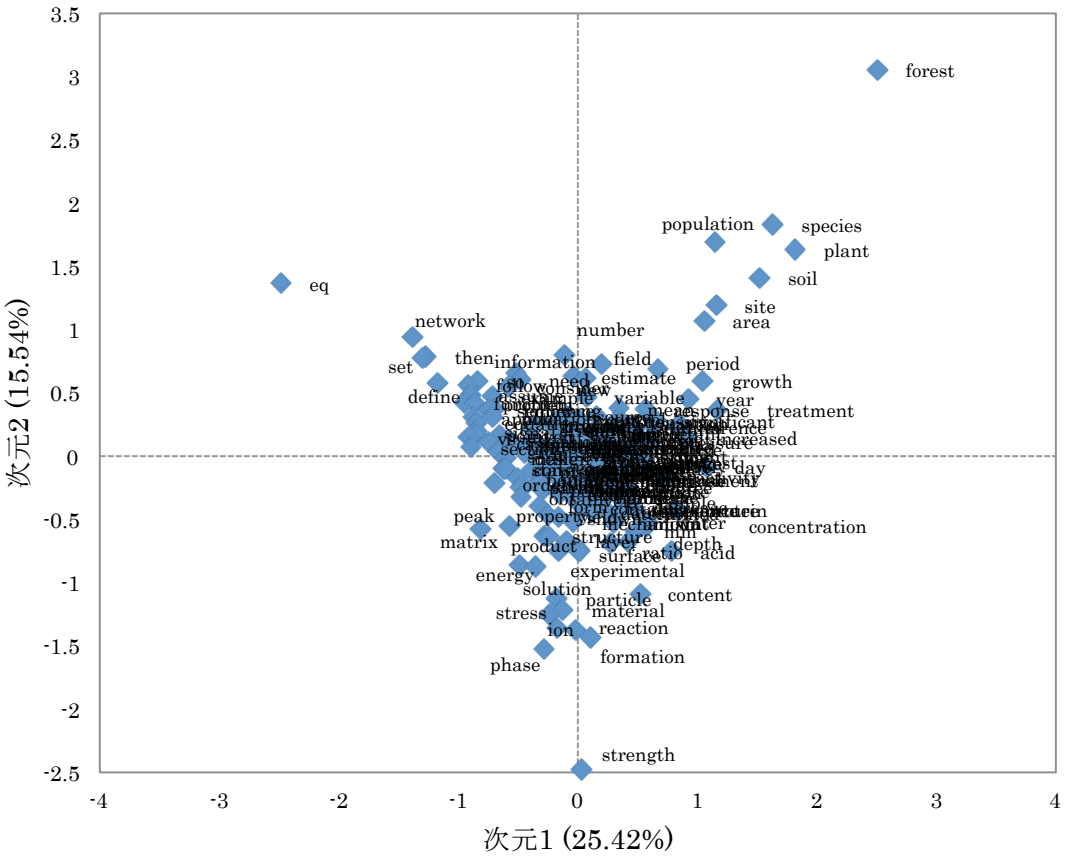

図 6　コレスポンデンス分析の結果（サンプル・スコア）

次元 1 は，図 4 の主成分分析の第 2 主成分と対応していると考えられるため，プラスが生物系分野を示し，マイナスが情報工学に関連した情報を示していると解釈できる。主成分分析の主成分と同じように，コレスポンデンス分析でも，各次元は互いに直交している（無相関である）ため（田畑，2004），次元 1 とは関係のない軸の解釈を次元 2 に対してしなければならない。ここでは，図 6 で示されたサンプル・スコアがマイナスの語には，energy, content, particle, material, stress, ion, reaction, formation, phase, strength のようなものがあり，プラスの語では，forest, species, population, plant, soil, eq, site, area, network, number のような語であるため，実験が行われる研究領域であるか（マイナスの値），そうでないか（プラスの値）を示している可能性がある。

最後に，上記のコレスポンデンス分析の結果を (a) 3.3. で行われた主成分分析の「第 2 主成分と第 3 主成分」，(b) 田畑 (2004) で推奨されている，「行列データを転置して行う主成分分析」の 2 つの結果と比べてみることにする。君山 (2002) は，コレスポンデンスと主成分分析の違いについて，「コレスポンデンス分析が，各要素のプロフィールのみを取り出し，その類似関係を見ようとしているが，主成分分析は，すべてを含んだ分散（2 乗和）を用いて分析し，その第 1 主成分の固有ベクトルとして，標準的なプロフィールを推定していると解釈することができる。したがって，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。」(pp. 30-31)，と述べており，今回のコレスポンデンス分析の結果と，主成分分析の「第 2 主成分と第 3 主成分」がどれほど類似しているのかを比べてみる。また，田畑 (2004) は，Burrows (1989) 以降，テクスト間の差を効果的に見るためには「行列データを転置して行う主成分分析」が使われていると主張しており，こちらの結果も合わせて検討してみる。

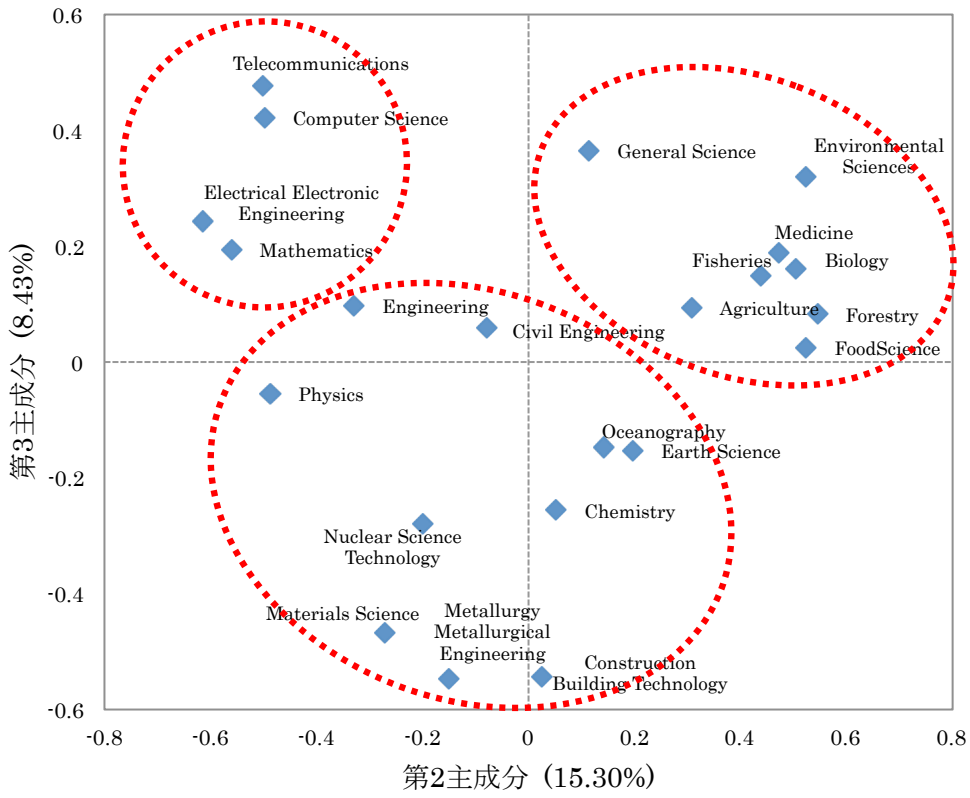

図 7　第 2 主成分と第 3 主成分をプロットしたもの

図 7 は，3.3. で行われた主成分分析の「第 2 主成分と第 3 主成分」をプロットしたものである（破線での領域分けはクラスター分析の結果に基づく）。図 5 のコレスポンデンス分析の結果と比べてみると，君山（2002）の言うように，領域のポジショニングは「第 2 主成分と第 3 主成分」と似ている結果になっている。また，コレスポンデンス分析で問題になった外れ値の影響が出ていない。

次に，「行列データを転置して行う主成分分析」の結果を図 8 に示した（破線での領域分けはクラスター分析の結果に基づく）。第 2 主成分のプラスとマイナスを逆転させれば，図 5 のコレスポンデンス分析と近い結果になっていることがわかるだろう。また，図 7 の「第 2 主成分と第 3 主成分」と比べた場合，寄与率が高いため，元のデータをよりよく反映しているといえる。

以上の結果から，(a) コレスポンデンス分析，(b) 主成分分析の第 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 3 つは，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ことが示唆された。今回の研究目的である，PERC コーパスの領域分類に関しては，これらの方法でどの領域が近いかを確認することは可能であるため，解釈に大きな違いが生じることはないと考えられるが，今回の結果だけでは限定的なため，この結果が再現されるかは，今後さらに検討していくべきであろう。

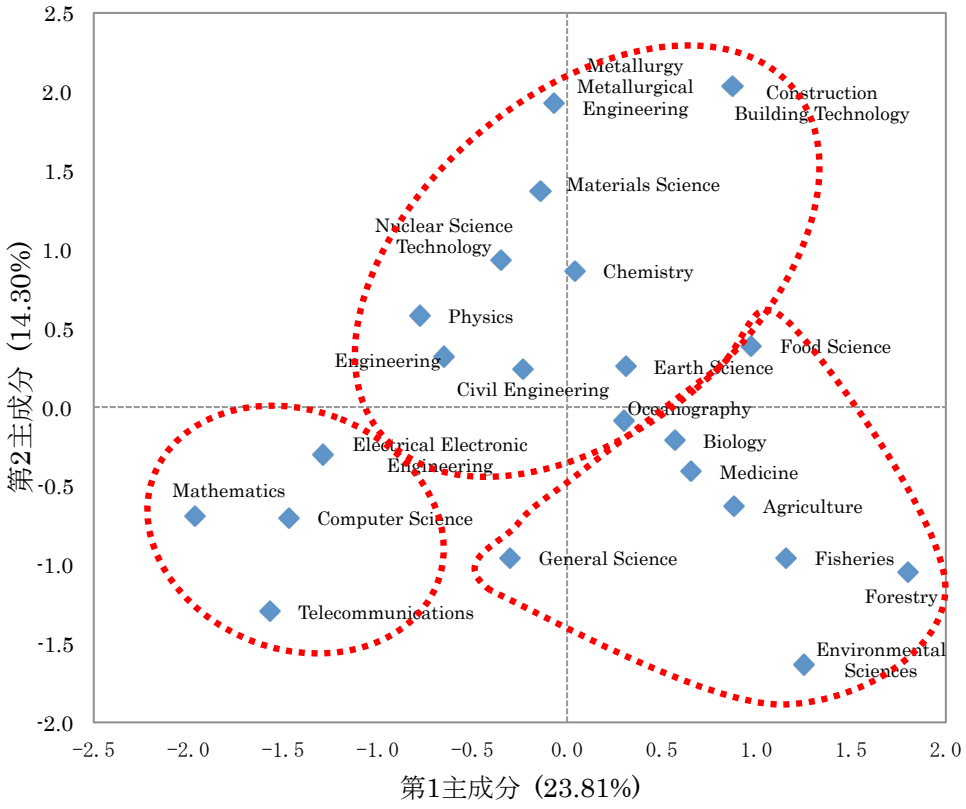

図 8　行列データを転置して主成分分析した結果

4. おわりに

本研究では，多変量解析を用いて PERC コーパスの領域がどのように分類可能であるかを検討することを目的とした。方法として，複数の変数間の関連性を検討する，圧縮・整理することが目的である，(a) クラスター分析，(b) 因子分析，(c) 主成分分析，(d) コレスポンデンス分析の 4 つを用いた。結果として，クラスター分析では領域ごとの近さや遠さを，因子分析では情報工学系分野，化学系分野，自然科学系分野，生物系分野という 4 つの潜在的な因子を抽出することができた。また，クラスター分析，主成分分析，コレスポンデンス分析では縮約されたデータにおける領域のポジショニングを，生物系，工学系，情報工学系の 3 つに大きく分類できることが確認された。

作成されたコーパスが妥当性の高いものであるかどうかは，この研究のように，当たり前であると思われる結果であっても，科学的なアプローチで証明していかなければならない。今回の研究の結果からは，PERC コーパスは科学技術英語コーパスとして，信頼の置ける言語資料であるといえることがわかった。また，コーパスの特性（学術論文のコーパスであること）からも，特に ESP（EAP）では利用価値が高いコーパスであると考えられる。今後の研究では，語彙レベルの分析だけでなく，コロケーションや lexical bundles のような，より大きな単位にも焦点を当て，PERC コーパスの分析を進めることにより，最終的には学習者が専門家集団（ディスコース・コミュニティー）の一員となるために必要な，より良い教材作成に PERC コーパスを利用する可能性を探っていくべきであろう。

謝　辞

本研究で使用した R のスクリプトは，法政大学の小林雄一郎氏に提供していただいたものです。ここに記して感謝いたします。

We would also like to thank Dr. Thomas Orr, PERC Board Chairman, Dr. Yukio Tono, Director of PERC, and Shogakukan for giving us access to the PERC corpus for research purposes.

文　献

朝野熙彦 (2007).『入門多変量解析の実際 [第 2 版]』東京：講談社サイエンティフィク.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: Benjamins.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26, 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at. . .: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25, 371-405.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*, 9-48.

Burrows, J. F. (1989). "'A Vision' as a Revision?" *Eighteenth-Century Studies*, *22*, 551-565.

Chen, Q., & Ge, G.-C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes, 26*, 502-514.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

大学英語教育学会基本語改訂委員会 (2003). JACET8000. 東京：JACET.

Dudley-Evans, T. (2001). English for specific purposes. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 131-136). Cambridge, UK: Cambridge University Press.

Dudley-Evans, T., & St John, M. J. (1998). *Developments in English for specific purposes: A multi-disciplinary approach*. Cambridge, UK: Cambridge University Press.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes, 19*, 115-135.

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.

Hyland, K. (2008a). Academic bundles: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18, 41-62.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, 4-21.

市川雅教 (2008).「Q62 被調査者の人数」繁桝算男・柳井晴夫・森 敏明 (編著)『Q & A で知る統計データ解析 DOs and DON'Ts [第 2 版]』(p. 125). 東京:サイエンス社.

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.

石川慎一郎 (2005).「司法英語 ESP 語彙表構築の試み―FROWN コーパスと米国司法文献コーパスの比較に基づく特徴語の抽出―」『神戸大学国際コミュニケーションセンター論集』, *1*, 13-28. Retrieved January 24, 2009, from http://iskwshin.googlepages.com/20050325_2.pdf

石川慎一郎 (2007).「コーパス高頻度語データにおける頻度分布の切断に対する相関行列の頑健性について―多変量解析を用いた言語コーパスのポジショニングを例として―」『統計数理研究所共同研究レポート No. 199 日英語の基本語抽出における統計手法の研究』 49-60. Retrieved January 24, 2009, from http://iskwshin.googlepages.com/20070315_2.pdf

磯田貴道 (2004). 「生徒のプロファイリング」 前田啓朗・山森光陽 (編著) 磯田貴道・廣森友人 (著)『英語教師のための教育データ分析入門：授業が変わるテスト・評価・研究』(pp. 112-124). 東京：大修館.

神宮英夫・土田昌司 (2008).『わかる・使える多変量解析』京都：ナカニシヤ出版.

川本竜史 (2004).『SPSS と Excel による統計力トレーニング』東京：東京図書.

菅 民郎 (2007).『Excel で学ぶ多変量解析入門 [第 2 版]』 東京：オーム社.

君山由良 (2002).『コレスポンデ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.

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.

小林雄一郎 (2007). 「The NICT JLE Corpus と語彙研究－SST レベルの再検証」『英文學誌』, *49*, 17-29. Retrieved January 24, 2009, from http://www.geocities.jp/cabinet_of_wonder/Eibungakushi49.pdf

古谷野 亘 (1988).『多変量解析ガイド』東京：川島書店.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

前田啓朗 (2004). 「自己評価項目の集約と解釈」 前田啓朗・山森光陽 (編著) 磯田貴道・廣森友人 (著)『英語教師のための教育データ分析入門 : 授業が変わるテスト・評価・研究』(第 10 章). 東京：大修館書店.

Ming-Tzu, K. W., & Nation, P. (2004). Word meaning in academic English: Homography in the Academic Word List. *Applied Linguistics, 25*, 291-314.

深山晶子 (編) (2000).『ESP の理論と実践：これで日本の英語教育が変わる』 東京：三修社.

村上征勝 (1994).『真贋の科学』東京：朝倉書店.

村上征勝 (2004).『シェークスピアは誰ですか？計量文献学の世界』東京：文藝春秋.
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.
Nakamura, J. (1995). Text typology and corpus: A critical review of Biber's methodology. *English Corpus Studies, 2*, 75-90.
小椋將弘 (2006).『Excel で簡単多変量解析』東京：講談社.
小塩真司 (2004).『SPSS と Amos による心理・調査データ解析―因子分析・共分散構造分析まで』 東京：東京図書.
Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th international ed.). Boston, MA: Pearson/Allyn & Bacon.
Tabata, T. (1995). Narrative style and the frequencies of very common words: A corpus based approach to Dickens's first person and third person narratives. *English Corpus Studies*, *2*, 91-109.
Tabata, T. (2002). Investigating stylistic variation in Dickens through correspondence analysis of word-class distribution. In T. Saito, J. Nakamura & S. Yamazaki (Eds.), *English corpus linguistics in Japan* (pp. 165-182). Amsterdam: Rodopi.
田畑智司 (2004). 「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 (2005).「コーパスに基づく文体論研究」 齊藤俊雄・中村純作・赤野一郎 (編)『英語コーパス言語学 基礎と実践 [改訂新版]』(pp. 183-206). 東京：研究社.
田地野 彰 (2004).「日本における大学英語教育の目的と目標について―ESP 研究からの示唆 ― 」 MM NEWS, 7. Retrieved January 24, 2009, from http://www.momiji.h.kyoto-u.ac.jp/MMpage/MM/MM7/MM7tajino.pdf
高橋 信 (2005).『Excel で学ぶコレスポンデンス分析』東京：オーム社.
対馬栄輝 (2008). 『SPSS で学ぶ医療系多変量データ解析』東京：東京図書.
Vongpumivitch, V., Huang, J.-Y., & Chang, Y.-C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes, 28*, 33-41.
安本美典・本多正久 (1981).『因子分析法』東京：培風館.