より良い学習語彙表の開発にむけた統計的手法の検討

その他のタイトル	Applying Statistical Methods to Develop a Better Vocabulary List
著者	水本 篤
雑誌名	統計数理研究所共同研究リポート199　「日英語の基本語抽出における統計手法の研究」
ページ	1-14
発行年	2007
URL	http://hdl.handle.net/10112/12984

統計数理研究所共同研究リポート 199
「日英語の基本語抽出における統計手法の研究」(2007), 1-14.

より良い学習語彙表の開発にむけた統計的手法の検討

水本　篤
大阪工業大学 非常勤講師
E-mail:　atsushi@mizumot.com

あらまし　これまでに数多くの学習語彙表が開発されているが，より良い語彙表を作るために，既存の語彙表における情報を統計的手法によって統合しているような研究はほとんどない。本研究では，基礎語の抽出のために，いくつかの語彙表などにおける語彙の特徴指標を統合し，より良い学習語彙表を開発するための統計的手法の検討を行うことを目的とした。まず，中学校主要英語教科書をコーパス化したものの頻度を基にした語彙表を作成し，いくつかの既存の語彙表や語に対する親密度，語の長さなどを語の特徴指標とし，それらをクラスター分析によってどのように分類できるのか確認し，因子分析によって潜在的な因子を調査した。いくつかの指標を統合するためには主成分分析が試された。結果として，これらの手法により，より良い学習語彙表が開発できるのではないかという可能性が示唆された。
キーワード　教育語彙表，語彙の特徴指標，多変量解析

Applying Statistical Methods to Develop a Better Vocabulary List

Atsushi, MIZUMOTO
Osaka Institute of Technology (Part-time lecturer)

Abstract　Many word lists for educational purposes have been developed thus far. However, no attempt has been made to integrate the information on the other word lists in the past with a statistical approach. The purpose of this study was thus to suggest statistical methods to incorporate several measures of individual words in order to develop a better basic word list. To this end, a junior high school textbook corpus was utilized to make a frequency-based word list out of it. The characteristics indices of vocabulary in this study include figures such as ranks or word levels in other word lists, word familiarity, and word length. With cluster analysis and factor analysis, categorization and latent factors behind these characteristics were examined. For the purpose of integrating the several measures for each word, principal component analysis was employed. The results suggest that the methods proposed here are promising in developing a better vocabulary list.
Keyword　Vocabulary Lists, Characteristics indices of vocabulary, Multivariate statistics

1. はじめに

1.1. 教育語彙表の重要性

語彙の習得が外国語学習者にとって最も重要なことの 1 つであることは，語彙学習が 4 技能（リスニング，リーディング，ライティング，スピーキング）に含まれていないにも

関わらず，語彙学習（語彙習得理論）に関する専門書（例えば Nation, 2001; 望月 他, 2003 など）が多いことからも容易に理解できる 。特に日本のように日常生活において英語に触れる機会が少ない EFL (English as a Foreign Language) の環境においては，欧米の研究で重要であるとされている，多読などによる偶発的語彙学習 (implicit vocabulary learning) が起こりにくい状況であると想像できる。日本の EFL 学習者たちにとって，中学，高校と英語を学ぶにあたって，最も大きな動機づけとなるのはテストであり，そのテストを受験する際には，学習者が知っている語彙と，テストに出題される語彙のサイズにギャップがあるという報告もあるため（Chujo & Nishigaki, 2003），学習者は語彙表（語彙リスト，もしくは一般的には「単語帳（集）」として知られているものを，そのギャップを埋めるためのツールとして用いることが多いと考えられる。また，基礎語については，文部科学省（2001）が中学校で学習すべき語を 900 語程度と決めていることからも，教科書でもある程度の語彙のレベル調整を行っていると思われる。このような理由から，どのような語を基礎語，もしくは学習すべき語として学ぶかは，教育語彙表に基づいていることが多いため，語彙表の開発は見過ごされがちではあるが，英語教育において大変重要な位置を占めているといえる。

1.2. 教育語彙表作成の歴史

教育的効果に重点を置いた語彙表は，外国語教育学が学問分野として確立していく過程と並行して開発されてきた。表 1 はそれらのうち，国内，国外において開発の分水嶺となったといえる語彙表まとめたものである（詳しくは，石川, 2007 を参照）。

表 1　過去における主要な教育語彙表

語彙表名	作成者	作成年
The Teacher's Word Book	Thorndike	1921
General Service List	West	1953
北海道大学英語語彙表	北海道大学言語文化部	1995
British National Corpus Frequency List	Kilgariff	1996
Academic Word List	Coxhead	2000
Standard Vocabulary List 12000	アルク	2001
JACET8000	JACET	2003

まず，Thorndike や West の時代には，電子データ化された大規模コーパスから語彙表を作成することができなかったが, それにも関わらず Thorndike の The Teacher's Word Book は 400 万語の資料から 1 万語の語彙表を作成したものであり，現在のコーパス言語学においても，このような語彙の抽出方法は行われている。General Service List は英語において

最頻出の 2000 語をまとめた語彙表であるというだけではなく，学習効果も考慮に入れて作られており，Aizawa（1998）で指摘されているように，リスト自体の古さなどの批判はあるものの，カバー率は現在においても高いことで知られている。また，Nation の Vocabulary Levels Test (1990)でも用いられている。1964 年に Brown Corpus が完成して以来，コンピューター・テクノロジーの発展とともに言語コーパスは次第に大規模になっていく。その流れ中で教育語彙表も大規模コーパスを反映したものが作成されるようになっていった。北海道大学英語語彙表は約 1200 万語の様々なコーパスの頻度と既存の 12 種類の教育語彙表を比較することによって作成された。このような語彙選定手法はコーパスの頻度のみに依存しないものとして注目に値する。1994 年に総語数 1 億語の British National Corpus（以下，BNC）が完成し，BNC における高頻度の語彙をまとめたものが Kilgariff の British National Corpus Frequency List である。また，Academic Word List（以下，AWL）は，学術誌と arts, commerce, law, natural science の 4 分野のサブ・コーパスを中心とする，大学のテキストの 350 万語から構成されているコーパスを基にして作られたものである。また，語彙の選定には頻度（frequency）とともに，「様々なアカデミック・テキストにどれだけ出てくるか」というレンジ（range）の考え方も用いられており，アカデミックな英文を読んだり書いたりするときに必須となる語彙を集めた語彙リストであるといえる。最後の Standard Vocabulary List 12000（以下，SVL12000）と JACET8000 は，両方とも BNC の頻度データを基にして作成されたものであるが，その語彙選択アプローチに違いがある。まず，SVL12000 は BNC の頻度とネイティブスピーカーの判断，そして日本語の使用頻度を考慮に入れることによって，主観を入れた語彙表となっており，一方，JACET8000 はサブ・コーパスなどを利用し，対数尤度比による頻度順位の補正を行うという客観的な手法を用いて作成された。

1.3. 特徴語抽出の研究

1.2.では包括的な語彙表の他に AWL が例として挙げられていたが，コーパス言語学の進歩は，EAP（English for Academic Purposes）や ESP（English for Specific Purposes），もしくは日本においては小学校英語などの特別な目的のための語彙表の開発にも繋がっている。通常，このような研究では基にするコーパス（例えば ESP コーパス）の頻度と，参照コーパス（BNC などの大規模な汎用コーパス）の頻度を，統計的指標を用いて比較することになる。中條・内山・長谷川（2005)の研究では，1）頻度，2）ダイス係数，3）補完類似度，4）対数尤度比，5）コサイン，6）イエーツの補正公式，7）カイ二乗値，8）自己相互情報量，という 8 種類の指標を比較し，頻度以外では，2）ダイス係数，3）補完類似度が初級レベル向けの語を選出し，4）対数尤度比，5）コサイン，6）イエーツの補正公式，7）カイ二乗値は中級レベル，そして，8）自己相互情報量は上級レベルという，8 つの統計指標が，学習者のレベルに合った語彙を選定できるということを示している。

石川（2005）は，KUBEE（Kobe University, Basic English Word Lists for Elementary

School Students）という日本人小学生向けの教育語彙表（約 1850 語）を開発した。この語彙表の開発では，語彙選定において頻度のみを重視することの問題点を補う視点として，児童のレベルにあった様々なコーパスと，児童がコミュニケーションをするときに必要である語を抽出するために，国語の教科書と児童の日本語作文で構成されている日本語コーパスを使い，インプット，アウトプットの両側面から望ましい語彙表が作成された。

これら２つの研究に共通していることは，１つのコーパスの頻度を基にした教育語彙表には限界があり，統計的手法や，学習者のレベルやニーズに合わせた視点を語彙表作成の際にも必要であるということであろう。

1.4. 語の指標を統合する手法を求めて(研究の目的)

語の特徴を示す指標を考えたときに，頻度以外に指標となるものを挙げると表 2 のようなものがある（中條 他, 2005; 池村, 2006; 石川, 2005; Laufer, 1997; Richards, 1970 などを参考に作成)。

表 2　語の指標

(1) 分布度（レンジ）
(2) 他のリストにおける順位（レベル）≒ 語の難易度・重要度
(3) 親密度
(4) 文字数（語の長さ）
(5) 多義性・具象性・抽象性
(6) 学習の容易さ（learnability）
(6) 音節数・発音の難しさ
(7) 派生語を取りうる数
(8) 連語度（collocability）
(9) L1 語彙としての使用（教科書や作文など）
(10) 書きことば・話しことばの違い（register）
(11) 受容語彙・発表語彙の違い
(12) 学習者のニーズ（language needs）

前述のような一連の研究で，教育語彙表は大規模コーパスやコンピューターを使うことによって，精緻化されてきたことがわかるが，さまざまな語彙リストや語の指標を統合するという観点の研究はこれまでになされていないといえるだろう。よって，本研究では，語彙表の精緻化を目指し，頻度を基本として作成される語彙表にさまざまな指標をプラスしていくことにより，どのような結果が得られるのかを検証することを目標とした。また，各指標がどのような関係にあるのかも同時に明らかにすることを試みた。

2. 方法

2.1. 使用したコーパスと語彙の選定方法

まず，基にするコーパスであるが，中学校の主要英語教科書をコーパス化したものを（検定教科書 3 冊，3 学年分）使用した。英語語彙の研究においては，何を 1 語と見なすかがよく問題となるので（例えば，look は名詞と動詞の用法なら 2 語と見なすのか，など），本研究ではJACET8000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，テキストの中で使われている語をレマ化し，1～8 までのランク付けとテキストカバー率を算出するものである。JACET8000 を基にしたランクの分類は表 3 のようになる。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）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 4 は，本研究で使用した中学校教科書コーパスを v8an で分析した結果である。延べ語数（tokens）は 13175 語となり，異なり語数（indexes）は 1444 語であった。この頻度を付与された語彙表の中から，省略語・固有名詞などの JACET8000 の 8000 語の中に入っていないものはカットし，頻度 1（＝レンジ 1）のものもカットした。また，後述する頻度以外の語の指標を表すデータベースに入っていなかった語を欠損値として削除して分析を行った。最終的に 541 語の語彙表となった。

表 3　v8an による語彙レベルの分類

分類（レベル）	説明
Level 1	ランクが 1 位～1000 位までの語
Level 2	ランクが 1001～2000 までの語
Level 3	ランクが 2001～3000 までの語
Level 4	ランクが 3001～4000 までの語
Level 5	ランクが 4001～5000 までの語
Level 6	ランクが 5001～6000 までの語
Level 7	ランクが 6001～7000 までの語
Level 8	ランクが 7001～8000 までの語
over 8	ランク外の語で，すべて小文字表記のもの（表 4 では over8）
cont. forms	シングルクォート（'）を含む語（表 4 では cont.）
proper nouns	大文字を含む語（表 4 では prop.）
non-words	アルファベット以外の文字を含む文字列（表 4 では non.）

表 4　中学校教科書コーパスの v8an による分析結果

	L 1	L 2	L 3	L 4	L 5	L 6	L 7	L 8	over8	cont.	prop.	non.	total
indexes	698	175	73	22	12	27	9	11	94	21	100	202	1444
tokens	9723	391	174	80	26	46	15	34	1683	374	148	481	13175

2.2. 語彙指標

2.1 で作成した 541 語の語彙表に対して，基本語彙性を表す指標を付与し，図 1 のようなデータセットが作成された。

今回使用された指標は以下の 9 種類だった。

1) 頻度（教科書コーパスのもの，B 列，変数名“Freq”），

2)レンジ（教科書コーパスのもので 9 が最大値，C 列，変数名“Range”），

3) JACET8000 の順位（8001 から順位の値を引くことで，数字の大きいものが基礎語であると考えられるように変換した，D 列，変数名“JACET8000”），

4) KUBEE1850 ver. 1.0.3（1847 語の語彙表なので，1848 から順位を引いた値，E 列，変数名“KUBEE”），

5) SVL12000 のレベル（JACET8000 と同じく順位を逆転させたもの，F 列）

6) BNC における頻度（G 列，変数名“BNC_Freq”）

7) 親密度調査から得られた評定（横川 他, 2006 の結果に基づく，1～7 の範囲，H 列，変数名“Familiarity”),

8）語の長さを数値化したもの（I 列，変数名“Length”），

9) 品詞の数（語の意味に含まれる品詞の数，JACET8000 の品詞タグに基づく，J 列，変数名“POS_Type”）

その他，1.4 で提案しているような，具象（抽象）性・類似性（音・意味），音節数・発音，派生語の数，連語度（collocability）などを数値でデータベース化したものは，今回は利用できなかった。

7) 親密度調査から得られた評定（横川 他, 2006）は，BNC の出現頻度上位 3000 語を対象として，日本人大学生約 800 名を被験者として「文字を見聞きする度合い」を調査したものである。また，8）の語の長さであるが，語の長さが認知レベルの上昇とともに長くなる傾向があると報告している研究もあるため（竹蓋 他, 1994），本研究でも指標として採用した。また，英語の多義性は語彙の学習においても学習者が困難に感じる側面であるため（投野, 1997），今回は必ずしも多義性を反映するものではないが，「その語の持つ品詞の数」を代理指標として入れた。例えば，have は動詞と助動詞の 2 種類の品詞があるので，指標は 2 となる。

	A	B	C	D	E	F	G	H	I	J
1	TheWord	Freq	Range	JACET8000	SVL12000	KUBEE	BNC_freq	Familiarity	Length	POS_type
2	be	506	8	13	1000	37	4239632	6.13	2	2
3	the	374	9	1	1000	1	6187267	6.79	3	1
4	I	364	9	7	1000	3	884599	6.82	1	1
5	you	302	9	10	1000	6	695498	6.92	3	1
6	a	283	9	5	1000	5	2186369	6.35	1	1
7	it	271	9	9	1000	8	1090186	6.64	2	1
8	in	236	9	6	1000	10	1924315	6.51	2	2
9	to	228	9	3	1000	4	1620850	6.76	2	3
10	do	224	9	20	1000	12	559596	6.41	2	2
11	have	151	9	15	1000	19	1375636	6.43	4	2
12	not	136	9	14	1000	35	465486	6.65	3	1
13	we	133	9	18	1000	33	358039	6.6	2	1
14	and	128	9	2	1000	2	2687863	6.8	3	1
15	this	125	9	25	1000	18	461945	6.23	4	2
16	of	113	9	4	1000	7	3093444	6.41	2	1
17	they	105	9	17	1000	26	433441	6.6	4	1
18	that	95	9	8	1000	13	760399	6.56	4	3

図 1　語彙表に各種の指標を付与したもの

2.3. 分析手法

今回の研究では，グループ間の差を検定する t 検定や分散分析のようなものではなく，表 5 にまとめられているように，たくさんの変数から 1 つの変数を予測したり，多くの変数間の関連性を検討したりする多変量解析が分析手法として用いられた。

表 5　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	因子分析	量的データ	
	主成分分析		
	クラスター分析		
圧縮・整理する	数量化 III 類	質的データ	
	コレスポンデンス （対応）分析		

多変量解析の中でも，今回の分析では変数間の関係を調査し，すべての変数を合成して，新しい指標（による順位）を作成することが目的であるため，因子分析，クラスター分析，そして主成分分析が試された。これらの分析にはすべて SPSS14.0 が使用された。

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単語）であることが，分類の目的であるため多いが，変数（図 1 では横の指標）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）。

（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，9 つの指標を並べていて，どのような因子があり，どのように分けることができるかを明らかにするのが目的である。具体的には，まず，1）変数間の相関係数を計算し, 2）因子の抽出を行い, そして, 3）因子を解釈しやすいように軸の回転を行う。4）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）。

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。SPSS のディフォルトでは因子分析の中に含まれることから，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）。主成分分析は，最も説明力の高い第一主成分を抽出するように分析が行われ，第二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。ゆえに，第一主成分と第二主成分は無相関（$r = .0$）となるようになっている（田畑, 2004）。

特に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）。今回の分析では，扱っているデータがコーパスにおける頻度であったり，語彙表の順位（順位を逆転させた形）であったために，正規性や線型性は保たれるようなものではなかったが，因子分析，クラスター分析，主成分分析などの分析では量的（相関関係が算出できる）データであれば分析が可能であり（出村 他, 2004; Tabachnick & Fidell, 2006），またデータの縮約と記述が目的であるため，本研究における分析手法として用いた。また，指標において，JACET8000，KUBEE，SVL12000 は，もともと語彙表に順位やレベルを付したものなので，尺度としては順序尺度と考えられるため，因子分析や主成分分析を適応するのは理論上好ましくないかもしれないが，今回は研究の目的が手法の模索であったために，そのまま分析に用いた。

本研究における，具体的な分析手法として，指標間の関係については分類と因子構造の確認となるので，クラスター分析と因子分析が用いられた。また，語彙表に付与された各種の指標における情報の圧縮が目的であるため，主成分分析によって語彙表の精緻化を試みた。

3. 結果と考察

3.1. 語彙指標の関係

多変量解析では，分析に先立って変数間の相関行列が確認される。表 6 は今回の研究で使用した各指標の相関行列を示している。

表 6　各指標の相関行列

	Freq	Range	JACET 8000	SVL 12000	KUBEE	BNC	Familiarity	Length	POS
Freq	1								
Range	.36**	1							
JACET8000	.28**	.52**	1						
SVL12000	.12**	.46**	.42**	1					
KUBEE	.32**	.73**	.63**	.51**	1				
BNC	.80**	.25**	-.23**	-.09*	-.24**	1			
Familiarity	.23**	.60**	-.52**	-.46**	-.52**	.15**	1		
Length	-.32**	-.34**	.29**	.18**	.39**	-.26**	-.15**	1	
POS	.05	.11*	-.15**	-.03	-.11**	.02	.07	-.19**	1

**$p < .01$, *$p < .05$

相関行列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，全体的にどの指標とどの指標が近い関係にあるのかを図示するために，それぞれの指標における数値を標準化した後に，クラスター分析（最遠隣法，ピアソン相関）を行った結果が図 2 である。

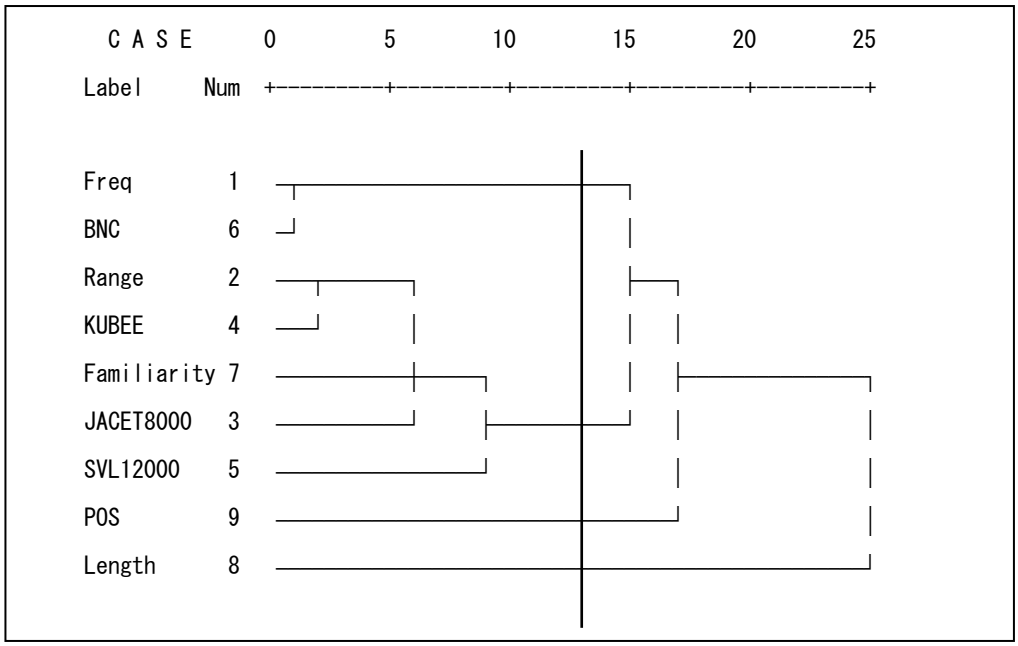

図 2　クラスター分析の結果（最遠隣法，ピアソン相関）

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，結合距離が 10 を超えた辺りで横の線が長くなっているため，中学校教科書コーパスの頻度（Freq）と BNC における頻度（BNC）のクラスターと，Range，KUBEE，Familiarity（親密度），JACET8000，SVL12000 のクラスターに分けることができる。POS（品詞の数）と Length（語の長さ）はこれらのクラスターには含まれるものではないため，他のクラスターとは離れている位置にある（関係があまりない）ことがわかる。

クラスター分析は変数の相関を用いて近いものをまとめていく手法であるため，因子分析でも同じような傾向が現れることが予想される。表 7 は各指標に対して，主因子法・プロマックス回転（因子の間に相関があることを仮定する回転）を用いて因子分析を行った結果である。表からもわかるように，Familiarity（親密度），Range（レンジ），KUBEE（小学生向け語彙表），SVL12000，JACET8000 が 1 つめの因子を構成し，Freq（中学生教科書コーパスの頻度）と BNC（BNC における頻度）が 2 つめの因子を構成している結果となった。これらはそれぞれ，語の難易度（Familiarity，Range，KUBEE，SVL12000，JACET8000）を反映している因子と，語の頻度（Freq，BNC）を反映している因子であると判断され，クラスター分析と同じ結果になっていることが確認できる。POS（語の品詞数）は十分な因子負荷量を示さなかったので，POS を除外して再度，因子分析を行ったところ（表 8），Length（語の長さ）は語の難易度の因子（Familiarity，Range，KUBEE，SVL12000，JACET8000）に含まれ，2 因子での累積寄与率が 65.3%になったが，因子負荷量が低かったため（ - .29），因子構造に含まれるべきではないと判断した。また，2 つの因子において，第一因子と第二因子の相関は $r = .37$ であり，それほど高い相関ではなかった。語の難易度の情報は，それぞれの語彙表の作成段階では何らかのコーパスの頻度に基づいているものの，語の頻度とはそれほど強い相関関係にないことが明らかになった。

表 7　因子分析の結果（3 因子）

	因子		
	1	2	3
Familiarity	.88	.03	-.26
Range	.75	.05	.08
KUBEE	.71	-.03	.25
SVL	.67	-.09	-.06
JACET8000	.63	.02	.10
Freq	.02	.93	-.01
BNC	-.04	.89	-.04
Length	.08	-.05	-.69
POS	-.03	-.09	.32

表 8　因子分析の結果（2 因子）

	因子	
	1	2
KUBEE	.85	.03
Range	.80	.06
Familiarity	.71	-.05
JACET8000	.69	.04
SVL	.64	-.12
Length	-.29	-.23
Freq	.01	.95
BNC	-.06	.85

3.2. 語彙表の精緻化

語彙表の精緻化のために，各指標を統合する目的で主成分分析を行った（表 9)。主成分の抽出には，相関行列による方法と，分散共分散行列による方法があるが，今回の分析では，各指標の単位が違うために（例えば，語の頻度と JACET8000 における順位など)，相関行列による方法で主成分分析を行った（分散共分散行列は単位が同じときに使われる)。

表 9　主成分分析の結果（1 回目）

	成分		
	1	2	3
KUBEE	.84	.19	.00
Range	.83	.17	.05
JACET8000	.75	.19	-.03
Familiarity	.71	.32	.18
SVL	.61	.39	.17
Length	-.51	.22	.44
BNC	.49	-.78	.14
Freq	.58	-.73	.11
POS	.18	.03	-.88

表 9 の結果では，POS（品詞の数）があるために第三主成分が抽出されていることがわかる。そのため，POS を除外して，もう一度主成分分析を行った結果が表 10 である。全分散のうち，2 つの主成分で説明される部分は 65.31%であった。表 10 から，第一主成分にすべての指標から重みがかかっていることがわかる。この結果から，第一主成分を「語の指標を統合したもの」と解釈することができる。Length（語の長さ）がマイナスとなっているのは，他の指標に比べて，語の長さが長くなるほど，難易度が高くなり，頻度も低くなるということを間接的に証明していると考えられる。

表 10　主成分分析の結果（2 回目）

	成分	
	1	2
KUBEE	.84	.20
Range	.83	.17
JACET8000	.75	.19
Familiarity	.71	.32
SVL	.61	.40
Length	-.50	.22
BNC	.50	-.77
Freq	.59	-.72

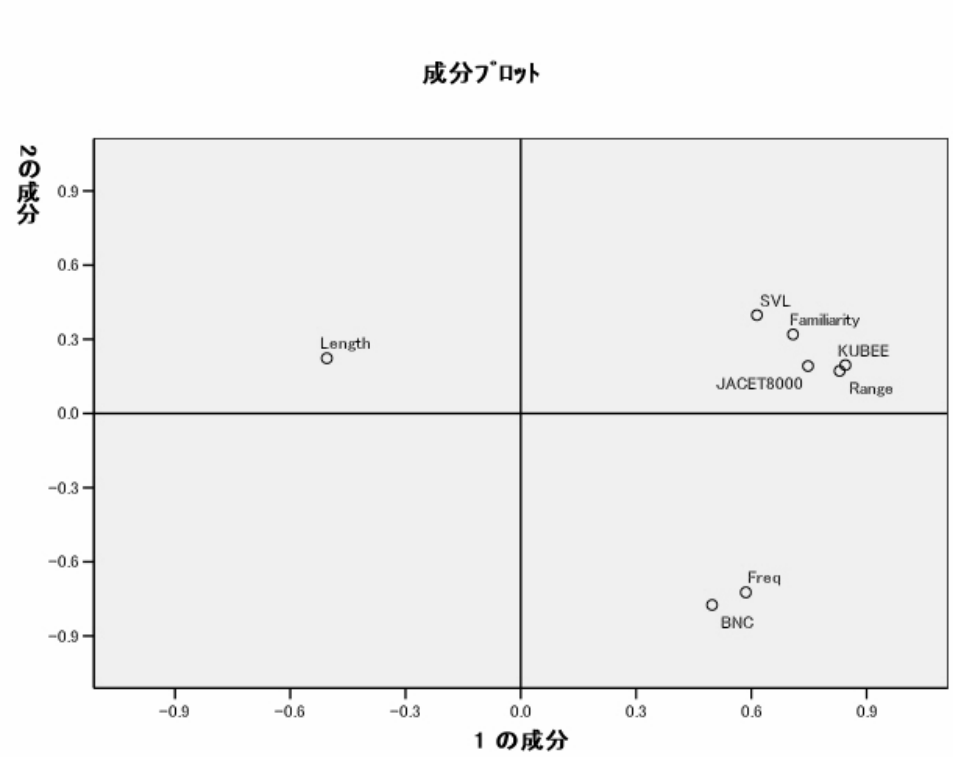

図 3　主成分分析の成分をプロットしたもの

図 3 は主成分分析の成分をプロットしたものである。横軸が第一主成分軸となる。この図から，クラスター分析の分類と因子分析による因子構造の確認における結果と同じ結果が得られたことがわかる。

最後に，主成分分析の結果得られた第一主成分得点と第二主成分得点を抽出し，中学生教科書コーパスの頻度のみで並べられた語彙表を，主成分得点を基にして並べ替えた。表 11 は第一主成分得点と第二主成分得点の相関係数をまとめたものである。頻度ベースの語彙表では同順位が多かったため，ケンドールの順位相関係数を求めたところ，相関係数は r = .62 であった。また，第二主成分との相関係数は r = .15 であった。第一主成分得点はすべての指標とある程度の相関関係にあるため，主成分得点によって並べ替えた語彙表は，語の特徴指標をまとめているものであると判断できる。一方，第二主成分得点では Range, SVL, Familiarity とのある程度の相関が見られた。第二主成分得点の順で並べ替えた語彙表の上位 20 語は順に，beautiful, yesterday, everything, question, different, station, morning, picture, after, little, something, house, important, understand, happy, difficult, teacher, sometimes, student, mother であった。これらの語には，機能語よりも内容語が多く含まれており，頻度を基にした語彙表とは違った語が抽出されていることが考えられる。

これらの結果から，主成分得点での並び替えにより頻度ベースの語彙表の順位が変動したことが確認され，第一主成分得点を利用して並べ替えたものは各指標を圧縮しているも

のであり，第二主成分を利用したものはそれ以外の要素をまとめている可能性があるということがわかった。

表 11　主成分得点と各指標の相関係数

	第一主成分得点	第二主成分得点
Freq	.62**	.15**
Range	.72**	.35**
JACET8000	.66**	.19**
KUBEE	.75**	.24**
SVL	.45**	.44**
BNC	.58**	.12**
Familiarity	.51**	.44**
Length	-.41**	.15**

**$p < .01$

4. おわりに

本研究では，より良い学習語彙表の開発にむけた統計的手法，特に多変量分析の適応の検討を行った。結果として，以下の 3 点が明らかになった。1）クラスター分析によって，どの指標が近いものであるかを明らかにすることができ，分類が可能である。2）因子分析によって，今回のように多くの指標が混在するときには因子構造を確かめることができる。更にその因子分析の結果は，クラスター分析の結果と符合するものであった。3）主成分分析で重み付けをした得点を利用し，単純な頻度とは異なる順位付けが可能であった。

今回の研究での結果の有効性を検証するために，より大きなコーパスを用いて，結果の再現性を検証することや，客観的な手法（統計的手法）で主成分分析を用いて，新しく並べ替えられた語彙表が本当に精緻化されているものなのか，主観（学習者や教師の評価）で判断するような調査を行う必要があり，今後の検討課題である。石川（2005）の目指した，学習語彙表における「客観と主観の融合」を実現するためには，更なる調査が必要である。

謝辞

本研究を行うにあたり，中学校教科書コーパスのデータをご提供いただきました，神戸大学の石川慎一郎助教授に深く感謝いたします。また，本稿について，同氏，及び統計数理研究所の前田忠彦助教授から，有益なコメントを多数いただきました。末筆ながらここに記して謝意を表します。

文　献

Aizawa, K. (1998). Developing a vocabulary size test for Japanese EFL learners. *Annual Review of English Language Education in Japan (ARELE) 9*, 75-85.

Chujo, K., & Nishigaki, C. (2003). Bridging the vocabulary gap: from EGP to EAP. *JACET Bulletin, 37*, 73-84.

中條清美・内山将夫・長谷川修治 (2005).「統計的指標を利用した時事英語資料の特徴語選定に関する研究」『英語コーパス研究』*12*, 19-35.

池村大一郎 (2006). 「既存データベースと単語親密度」横川博一 (編著)『日本人英語学習者の英単語親密度：文字編』(第 5 章). 東京：くろしお出版

石川慎一郎 (2005).「日本人児童用英語基本語彙表開発における頻度と認知度の問題：母語コーパスと対象語コーパスの頻度融合の手法」『信学技報（電子情報通信学会）』*25*, 43-48.

石川慎一郎 (2007).「英語教育のための基本語をどう選ぶか：コーパス言語学からの視点」『英語教育』*55*(13), pp. 10-13.

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』(第 12 章). 東京：大修館.

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

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

Laufer, B. (1997). What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 140-155). Cambridge: Cambridge University Press.

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

望月正道・相沢一美・投野由紀夫 (2003).『英語語彙の指導マニュアル』東京：大修館書店

Nation, P. (1990). *Teaching and learning vocabulary*. New York: Newbury House Harper Row.

Nation, P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Richards, J. C. (1970). A psycholinguistic measure of vocabulary selection. *IRAL*, *8*, 87-102.

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

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

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th international ed.). Boston, MA: Pearson/Allyn & Bacon.

竹蓋幸生・長谷川修治・中條清美 (1994).「語彙リスト：『現代英語のキーワード』の認知レベルによる区分の妥当性」『言語行動の研究』*4*, 53-63.

投野由紀夫 (編著) (1997).『英語語彙習得論 ボキャブラリー学習を科学する』東京：河源社

横川博一 (編著) (2006).『日本人英語学習者の英単語親密度：文字編』東京：くろしお出版