主成分分析を用いた学習語彙表の精緻化の試み

その他のタイトル	An Attempt to Elaborate a Vocabulary List with Principal Component Analysis
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	15-26
発行年	2007
URL	http://hdl.handle.net/10112/12983

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), 15-26.

主成分分析を用いた学習語彙表の精緻化の試み

水本　篤

大阪工業大学 非常勤講師

E-mail:　atsushi@mizumot.com

あらまし　本研究では, 水本（2007）で検討された主成分分析を用いた学習語彙表の精緻化を試みた。はじめに，水本（2007）よりもサイズの大きいコーパス（中学校検定教科書 7 冊 3 学年分）を用い，同じ手法で結果を検証することにより，提案した手法に再現性があるのかを検証した。また，主成分分析によって新しく並べ替えられた語彙表が，教育的にも有効であり，精緻化されていることを実証するために, 教育現場で指導に携わっている英語教員に主観による判断を求めるアンケート調査を実施した。その結果，主成分分析を用いた精緻化の方法は，基礎語レベルにおいてはそれほど有効ではなく, 従来の頻度ベースの語彙表の方が基礎語の学習には重要である可能性が示唆された。

キーワード　語彙表，主成分分析，アンケート調査

An Attempt to Elaborate a Vocabulary List with Principal Component Analysis

Atsushi, MIZUMOTO

Osaka Institute of Technology (Part-time lecturer)

Abstract　In this paper, a new method suggested by Mizumoto (2007) for elaborating word lists for educational purposes is verified. First, a large corpus composed of junior high school textbooks was used to validate the results in the preliminary study. In addition, the elaboration of the word list with the help of principal component analysis was tested by administering a questionnaire to the professional teachers in service at different stages of institutions. They decided whether they felt the list devised was worth teaching to their students. The results indicate that the newly developed method for elaborating a word list may not be useful in distinguishing basic words compared to the de facto standard－a frequency-based list. Several reasons are explored to account for the unexpected result.

Keyword　Word lists, Principal component analysis, Questionnaire survey

1. はじめに

水本（2007）の研究においてはいくつかの語彙の特徴指標をクラスター分析で分類し, 因子分析で潜在因子を探り，主成分分析を用いて最終的に語彙表を精緻化する方法として提案しているが，本稿ではその手法をより大きなサイズのコーパスを用い，同じ手法で結果を

検証することにより，提案した手法に再現性があるのかを検証した。また，主成分分析によって新しく並べ替えられた語彙表が，教育的にも有効であり，実際に精緻化されているのかということを検討するために，教育現場で指導に携わっている英語教員に主観による判断を求めるアンケート調査を実施した。

2. 方法

2.1. 基データ

水本（2007）の研究では，中学校教科書コーパスを検定教科書 3 冊で 3 学年分（延べ語数 13175 語）のもの使用したが，今回の研究では，中村（2006）で作成された検定教科書 7 冊の 3 学年分の教科書コーパス（延べ語数 112879 語）から抽出され，頻度が付けられた語彙表を使用した。

まず，語彙表における語のカウント方法が中村（2006）と水本（2007）の研究では違ったので，水本（2007）と同じ基準でレマ化を行った。つまり，JACET8000 に合わせる形を基本とし，@. &. ?などの記号が削除され，‘s は所有格，has, is などさまざまな解釈の可能性があるため，今回の分析では排除された。また，短縮形はそれぞれの語を'll は will，’re と’m は be，’ve は have という形で変換し，それぞれの頻度に加えた。Dr. Ms. Mr. TV, CD などの語は基礎語であるのは確かであるが，今回はカットした。そして，数字なども語彙表から除き，最後に水本（2007）の研究と同じように，頻度 1（レンジ 1）のものをカットし，頻度以外の語の指標を表すデータベースに該当する語が入っていなかった場合には，欠損値として削除して分析を行った。これらの基準に沿って語を選定していくと，最終的に 891 語の語彙表となった。この語彙表を基に，水本（2007）の研究と同じ分析方法を用い，結果の再現性の確認を行った。

2.2. 語彙指標

図 1 は語彙表に含まれた 891 語に対して，各指標を付与したものである。水本（2007）の研究と同じように，今回使用された指標は以下の 9 種類であった。また，表 1 は各指標の基礎統計となっている。表 1 の歪度に示されるように，Freq，BNC などコーパスにおける出現頻度に関わる指標は，右に強く歪んだ分布を持つ指標であるとわかる。

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 7 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ように変換した，D 列，変数名“JACET8000”），
4) SVL12000 のレベル（逆転させたもの，E 列，変数名“SVL12000”）
5) KUBEE1850 ver. 1.0.3（1847 語の語彙表なので，1848 から順位を引いた値，F 列，変数名“KUBEE”），
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”）

	A	B	C	D	E	F	G	H	I	J
1	TheWord	Freq	Range	JACET8000	SVL12000	KUBEE	BNC_freq	Familiarity	Length	POS_type
2	I	4022	7	7994	12	1845	884599	7	1	1
3	the	3587	7	8000	12	1847	6187267	7	3	1
4	you	2983	7	7991	12	1842	695498	7	3	1
5	be	2547	7	7988	12	1811	4239632	6	2	2
6	to	2395	7	7998	12	1844	1620850	7	2	3
7	a	2214	7	7996	12	1843	2186369	6	1	1
8	do	1940	7	7981	12	1836	559596	6	2	2
9	in	1661	7	7995	12	1838	1924315	7	2	2
10	not	1377	7	7987	12	1813	465486	7	3	1
11	it	1210	7	7992	12	1840	1090186	7	2	1
12	have	1156	7	7986	12	1829	1375636	6	4	2
13	and	1046	7	7999	12	1846	2687863	7	3	1
14	this	945	7	7976	12	1830	461945	6	4	2
15	he	874	7	7989	12	1833	681255	7	2	1
16	can	868	7	7965	12	1799	266116	6	3	3
17	we	850	7	7983	12	1815	358039	7	2	1
18	what	809	7	7972	12	1828	249466	7	4	1
19	of	807	7	7997	12	1841	3093444	6	2	1
20	for	797	7	7990	12	1827	887877	7	3	3

図 1　語彙表に各種の指標を付与したもの

表 1　各指標の記述統計量

	最小値	最大値	平均値	標準偏差	歪度	尖度
Freq	2	4022	88.13	293.53	8.29	84.33
Range	1	7	4.51	2.29	-0.23	-1.52
JACET8000	5559	8000	7363.96	518.99	-1.05	0.57
SVL12000	4	12	11.74	0.57	-4.34	40.87
KUBEE	19	1847	1214.41	453.88	-0.66	-0.45
BNC	1252	6187267	69076.26	321911.52	12.60	194.84
Familiarity	3	7	5.92	0.68	-1.28	2.03
Length	1	11	4.96	1.64	0.85	0.78
POS	1	6	1.42	0.62	1.73	5.28

2.3. リストの妥当性に関する検証方法

水本（2007）の研究と同じように本研究においても，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を用い，語彙表の精緻化は主成分分析によって行われた。また，統計手法を使って客観的に精緻化を行った語彙表が，主観的判断においてどれぐらい基礎語の抽出に役立っているのかを調査するために，現職（中学，高校，大学）の英語教員を対象にアンケート調査を行った。

3. 結果と考察

3.1. 語彙指標の関係

表 2 は今回の研究で使用した各指標の相関行列を示したものである。この相関係数の比較によって，それぞれの指標がどの程度の関係があるのかを確認することができる。まず，Freq（頻度）は BNC（BNC における頻度）とかなり高い相関係数となっている（$r = .75$）。そして Range（レンジ）は KUBEE（$r = .64$），そして Familiarity（$r = .61$）と.60 以上の相関係数を示している。同様に，JACET8000 や SVL12000 と他の指標を比べてみると，それぞれの指標と近いものがある程度確認できる。

表 2　各指標の相関行列

	Freq	Range	JACET 8000	SVL 12000	KUBEE	BNC	Familiarity	Length	POS
Freq	1								
Range	.29**	1							
JACET8000	.27**	.58**	1						
SVL12000	.13**	.47**	.39**	1					
KUBEE	.32**	.64**	.62**	.37**	1				
BNC	.75**	.19**	.22**	.08*	.23**	1			
Familiarity	.25**	.61**	.57**	.54**	.46**	.15**	1		
Length	-.27**	-.17**	-.18**	-.17**	-.30**	-.21**	-.08*	1	
POS	.06	.06	.19**	.07*	.12**	.02	.08*	-.18**	1

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

次に，それぞれの指標における数値を標準化し，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。結果として，水本（2007）と同じ傾向のクラスターに分類され，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，結合距離は水本（2007）の結果とは少しだけ違うものの，Range，KUBEE，Familiarity（親密度），JACET8000，SVL12000 のクラスターに分けることができることが明らかになった。また，POS（品詞の数）と Length（語の長さ）は他のクラスターと離れている位置にあることが図からわかる。この結果も水本（2007）で得られたものと同じであり，結果の一貫性から，今回使用している指標の内部構造はコーパスのサイズが大き

くなり，頻度やレンジの数値が変わったとしても同じものであると考えられる。

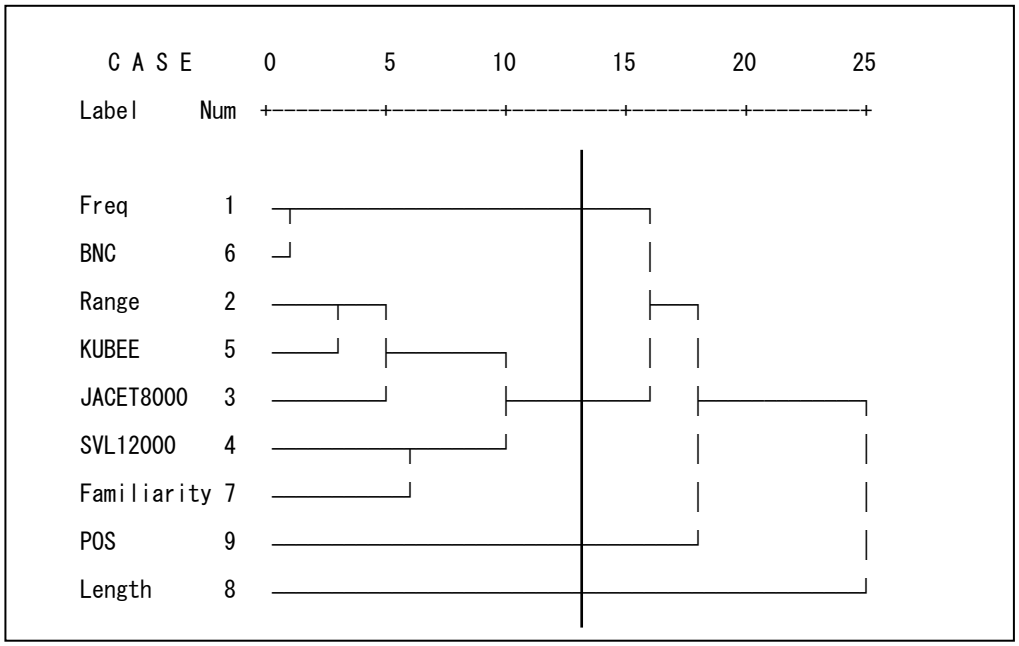

図 2　クラスター分析の結果（最遠隣法，ピアソン相関）

表 3 は各指標を用いて，主因子法・プロマックス回転を用いて因子分析を行った結果である。POS（語の品詞数）と Length（語の長さ）は水本（2007）の研究と同じように十分な因子負荷量を示さなかったので，これら 2 つの指標を除外して再度，因子分析を行ったところ，語の難易度の因子（Range，Familiarity，JACET8000，KUBEE，SVL12000）と，語の頻度の因子の 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と考えられる。二因子での累積寄与率は 69.9%になった。また，2 つの因子において，第一因子と第二因子の因子間相関は $r = .35$ であった。

表 3　因子分析の結果

	因子	
	1	2
Range	.81	.05
Familiarity	.77	-.04
JACET8000	.73	.04
KUBEE	.70	.09
SVL12000	.62	-.10
Freq	.04	.88
BNC	-.06	.87

3.2. 語彙表の精緻化

各指標を統合する目的で，相関行列による方法を用いて主成分分析を行った。その結果，水本（2007）と同じように，POS（品詞の数）があるために第三主成分が抽出されていたので，POS を除外して，もう一度主成分分析を行った。表 4 がその結果である。この表を見ると，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ので，第一主成分が「語の指標を統合したもの」と解釈することができる。

表 4　主成分分析の結果

	成分	
	1	2
Range	.81	.24
KUBEE	.79	.10
JACET8000	.77	.19
Familiarity	.75	.33
SVL12000	.62	.36
Length	-.37	.32
BNC	.47	-.77
Freq	.57	-.71

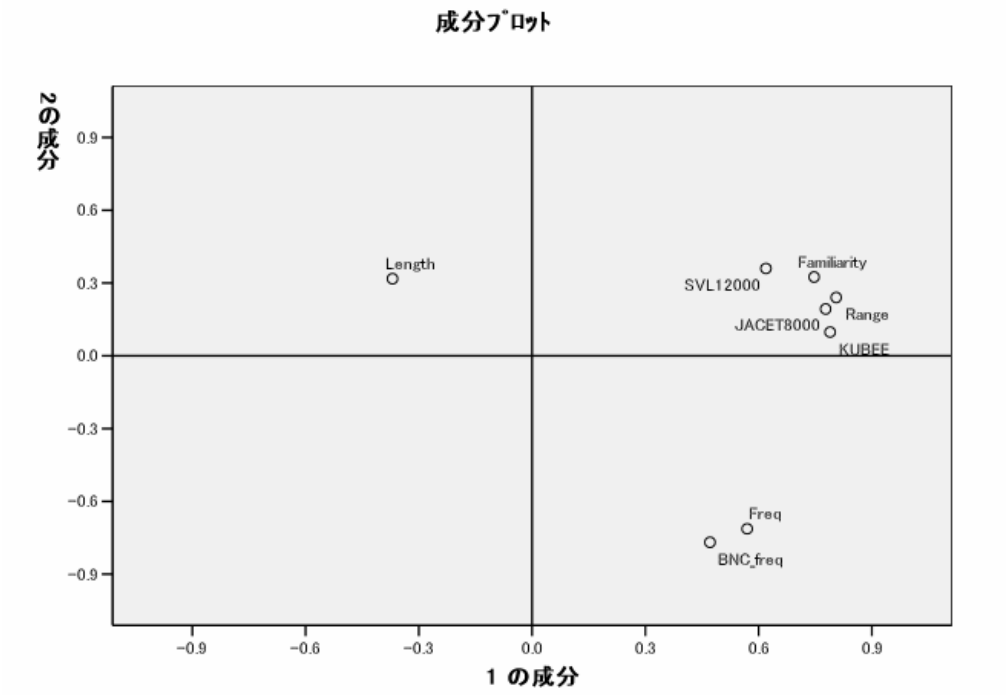

図 3　主成分分析の成分をプロットしたもの

図 3 は横軸が第一主成分軸として，主成分分析の成分をプロットしたものである。この図から，水本（2007）の研究において得られた結果と同じように，クラスター分析による分類と似ている結果が得られたことがわかる。

また，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と中学校教科書コーパスの頻度とのケンドールの順位相関係数を求めた（表 5）ところ，頻度との相関係数第一主成分が $r = .69$ であった。また，第二主成分との相関係数は $r = .32$ であった。水本（2007）の研究と同じように，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，語の特徴指標を圧縮しているものであると判断できる。一方，第二主成分得点の順で並べ替えた語彙表の上位 20 語は，everything, understand, different, beautiful, important, interesting, sometimes, difficult, question, something, yesterday, anything, restaurant, tomorrow, wonderful, remember, together, nothing, afternoon であった。第二主成分によって並べ替えられた語彙表も頻度以外の指標との相関がある程度見られることから，語彙表の精緻化という観点からは検討に値すると考えられるが，主成分分析の目的（第一主成分に情報を圧縮する）から考えても，今回の研究では第一主成分に基づいて順位を入れ替える方が妥当であると考えられた。

表 5　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	.69**	.32**
Range	.71**	.46**
JACET8000	.66**	.27**
SVL12000	.52**	.45**
KUBEE	.69**	.23**
BNC_freq	.58**	.21**
Familiarity	.57**	.44**
Length	-.28**	.25**

** $p < .01$

3.3. アンケート調査

語彙表における語の並べ替えは，891 語すべてに対して行われたが，現職英語教員に対する調査では，基礎語の抽出における妥当性を検証するため，上位 100 語のみを対象として，頻度ベースの語彙表にのみ出現する語 22 語を「リスト B」とし，第一主成分得点で並べ替えた語彙表にのみ出現する語 22 語を「リスト A」として，どちらのリストの方が基礎語として学習者に提示すべきものであるかを尋ねるアンケート調査を行った（アンケート調査の内容は巻末の資料を参照）。また，提示の順番による効果を相殺するため，半数の回答者にはリストの A と B を入れ替えた調査用紙を配布して，カウンターバランスを取った。100 位までで共通していた 78 語と，それぞれのリストにしか現れなかった 22 語は以下のような語となっている。

【100 位までで両語彙表に共通していた 78 語】

the, be, I, to, a, you, in, and, of, do, it, have, for, not, that, he, like, no, on, go, we, can, this, my, she, what, will, by, at, me, yes, they, with, but, how, play, up, time, all, as, good, very, there, from, his, who, your, want, when, make, about, her, right, day, year, take, see, use, people, look, help, why, some, book, school, many, so, now, well, get, know, home, last, come, much, our, here, think

【頻度ベースの 100 位以内にのみ出現していた 22 語】

thank, too, friend, one, please, where, live, student, old, hello, call, him, watch, them, lot, visit, really, tennis, nice, which, yesterday, eat

【主成分得点で選んだ語の 100 位以内にのみ出現していた 22 語】

if, out, over, long, say, back, just, only, after, off, other, try, new, more, name, work, man, talk, ask, give, open, house

アンケートはできるだけ広い層の英語教師の回答を得るために，中学，高校，大学（短期大学も含む）の英語教員 47 名（指導経験年数 $M = 11.15$，$SD = 7.9$）に調査を行った。表 6 はアンケート調査の結果を集計したものである。

表 6　アンケート調査の結果

	主成分分析を使ったリスト	頻度ベースのリスト	合計
中学	0	8	8
高校	1	24	25
大学	0	11	11
その他	0	3	3
合計	1	46	47

表 6 から明らかであるように，頻度のみの情報を基にした語彙表を調査した 47 名中 46 名が選んだという結果になった。頻度ベースのリストを選らんだ回答者のほとんどのコメントに，「コミュニカティブな語，会話で使われるが多く含まれている」や，「リスト A（主成分分析を使ったリスト）は if や after など機能語が含まれていて，どちらかといえばコミュニケーションに重要なのはリスト B だと思う」，「語の導入を行うときに，身近な語の方が関心を引くことができるから」などというものが含まれていた。

これらのコメントからも，いくつかの基本語彙指標であると考えられる指標を用いて，

主成分得点により語彙表の並べ替えを行ったものの，英語教師の主観による判断では頻度ベースの従来の語彙表の方がふさわしいという結果となった。

4. おわりに

今回の研究の結果についてまとめる前に，本研究におけるいくつかの研究デザインの問題点を指摘しておく。まず，いくつかの語の指標をコーパスから作成された語彙表に付与して分析を進めたが，その際に，各指標のデータベースに入っていなかった語を，欠損値があるために，分析の対象から外した。今後の研究では，このような欠損値の扱いをどうするべきか検討が必要である。そして，今回の研究で関係を調査した語の指標であるが，JACET8000，KUBEE は語彙表における順位，そして SVL12000 はレベルを指標の数値としたので，尺度としては順序尺度となり，因子分析，主成分分析で分析可能とされる量的データ（出村 他，2004）であるとは言い難い。このような統計手法を利用するときに注意すべき尺度に関する問題も，今回の研究で行ったような分析を今後適応する際には検討していくべきであろう。

上記のような問題点を踏まえつつ，今回の研究での結果をまとめると以下のようなものであった。

1) 水本（2007）の研究よりもコーパスサイズを大きくして，語の指標間の関係を調査したところ，サンプル数をかなり増やしても同じ結果になった。
2) 主成分分析によって統合された新しい語彙表は，基礎語である 100 語までは従来の頻度ベースの語彙表と比べると，英語教員の主観によるアンケート調査の結果から，頻度ベースの語彙表の方に基礎語が含まれているという結果になった。

より良い教育語彙表作成のために主成分分析を用いて統合指標を作り，その有効性を検証するのが本研究の目的であったが，結果は予想していたものとは逆のものとなった。本研究の（特にアンケート調査の）結果の原因として，以下の可能性が考えられる。

1 つ目の可能性は，「目的と研究手法が一致していなかった」ということである。今回の研究の目的は「基礎語の語彙表を精緻化すること」であったが，主成分分析で統合指標を作成する際に基となった指標は，JACET8000 や SVL12000 などのような，英語入門レベルの学習者を必ずしも対象としているものではなかったために，そのレベルにあった統合指標を作ることができなかったという可能性がある。特に，今回取り入れた各指標は受信型の情報が中心で，学習者が「こういう語を使って，こういう話がしたい」というような発信型の指標は石川（2005）で作成された KUBEE 以外に含まれていなかったために，このような結果となったのではないかとも考えられる。そして，クラスター分析や因子分析でも，語の難易度と語の頻度以外の指標がクラスターや因子として抽出されることがなかったために，語の learnability（学びやすさ）などの指標が，語の長さ以外になかったためにこのような結果になったという可能性も残されている。今後はこのような基礎語の特徴

をもっと反映していると考えられる指標を用いてみて，今回の結果と比較することも重要である。また，今回の研究では語彙表の並べ替えは主成分分析を行った際に得られた第一主成分得点を基に行ったが，第二主成分を用いて並べ替えを行ったものが，アンケートなどの主観で判断する調査を行えば，どのような結果が得られるのかという点については確認できていないため，この点については今後の検討課題であるといえよう。

2 つ目の可能性は，「基礎語レベルでは，教科書頻度の方がもともと重要であるため，このような統計的手法によって選ばれた語彙表よりも優れている」というものである。日本の EFL 環境において，中学生のインプットとして最も重要なものは教科書であり，その教科書を作成する際には，語彙のレベルなどの管理も行われているであろうと考えられる。ゆえに，その教科書をコーパス化して作成された頻度ベースの語彙表で，最も重要な基礎語を学ぶことができるようになっているという可能性がある。この解釈で考えると，英語教師の主観を調査したときに，頻度ベースの語彙表が圧倒的に選ばれたという事実も理解することができる。それゆえ，基礎語レベルの語彙表の精緻化ではなく，ESP（English for Specific Purposes）や EAP（English for Academic Purposes）などのある程度レベルの高い語彙表の精緻化の目的であれば，この手法が活きてくるのかもしれないために，それを検証していくことが必要になる。

主観だけではなく，統計に代表される客観的な手法を用いて，科学的なアプローチで重要な語彙を選定する研究は，コーパス言語学の発展のおかげでこれからますます進んでいくであろうと考えられる。しかし同時に，学習者が「どのように語彙を覚えるべき」か，また教師が「どう教えるべきか」という，「どのように」という側面に関する研究も進めていく必要があるといえる。

謝辞

本研究の実施にあたり，中学校英語検定教科書の語彙表データを立命館大学大学院言語教育情報研究科の中村純作教授にご提供いただきました。また，神戸大学の石川慎一郎助教授には，アンケート調査用紙の作成の際に貴重なアドバイスをいただき，本稿を作成するにあたり，同氏，及び統計数理研究所の前田忠彦助教授からご指導を賜りました。ここに記して心から感謝いたします。そして，お忙しい中アンケート調査にご協力いただいた先生方にお礼申し上げます。

文　献

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.

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
中村純作 (2006).「教科書コーパスから何が見えるか：方法論と中学校英語教科書の場合」『立命館言語文化研究』*17*(4), 143-166.
水本篤 (2007). 「より良い学習語彙表の開発にむけた統計的手法の検討」『日英語の基本語抽出における統計手法の研究』統計数理研究所共同研究リポート 199.
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

資　料　　現職の英語教員を対象とした基礎語語彙表の調査

指導されている教育機関　　中学校 / 高校 / 大学（短大も含む）/ その他
英語の指導を始められてから何年ですか？　（　　）年

Q. 英語の初学者（以前に全く勉強したことがない人）に示すべき最も重要な 100 語を選定するとします。そこに含めるものとして，次の語群のうち，リスト A かリスト B のいずれか 1 つだけを採用するとすれば，どちらがいいでしょうか？また，その理由はどうしてですか？

(A)	(B)
if	thank
out	too
over	friend
long	one
say	please
back	where
just	live
only	student
after	old
off	hello
other	call
try	him
new	watch
more	them
name	lot
work	visit
man	really
talk	tennis
ask	nice
give	which
open	yesterday
house	eat

どれかに○をつけてください
（　　）A の方が良い
（　　）B の方が良い

理由をお書き下さい

ご協力いただきましてありがとうございました。