統計的指標を利用した特徴語抽出に関する研究

Using Statistical Measures to Extract Specialized Vocabulary from a Corpus

中條清美
CHUJO Kiyomi
日本大学
Nihon University

内山将夫
UTIYAMA Masao
通信総合研究所
Communications Research Laboratory

Abstract

Earlier studies have established that the use of frequency of occurrence is effective in extracting specialized vocabulary from a corpus. What would happen if, rather than relying on solely frequency, a range of various statistical tools were used? In this study, eight individual and one 'F_{cum}' combination statistical analyses were evaluated for effectiveness in producing specialized vocabulary by comparing extracted lists to existing specialized vocabulary control lists. It was found that the 'F_{cum}' combination of measures created the most comparable data followed in effectiveness by the Dice coefficient. It was determined that all these measures were effective tools in producing beneficial specialized vocabulary, and that each measure created a unique list with regard to frequency, word length, type of word, and school textbook vocabulary coverage. While the use of frequency alone as a determiner of specialized vocabulary from a corpus is effective, the application of statistical tools provides even greater effectiveness in extracting various types of specialized lists which can be targeted to students' vocabulary or proficiency levels.

Key Word : 統計的指標　特徴語　語彙選定　自然言語処理

1. はじめに

英語力の基礎となる語彙は教科書や教材作成の際に重要な指針となるため，我が国でも全英連語彙集(1981), JACET 基本語 4000(1983), 東京都中英研(1986), キーワード 5000(1994), 北大語彙表（1996）など英語教師・研究者によってさまざまな語彙リストが作られてきた。最近のコーパス言語学の発展とともに，従来の頻度（frequency）や分布度（range）に選定者の主観的な知識，経験，勘を加えた語彙選定にも，自然言語処理の分野で利用されている統

計的指標が利用できるようになってきた。その一例として，70 名の英語教育関係者が関わって改訂された JACET8000（2003）では 1 億語の British National Corpus の頻度データをベースに, 日本人英語学習者に有用と思われる英文資料を照合して, 対数尤度比（log-likelihood ratio）という統計値に基づいて，単語の頻度順位を補正するという手法が用いられている。

対数尤度比のような統計的指標は単語どうしの共起頻度を利用して類似度を計算するもので，代表的なものに自己相互情報量(MI-score)，カイ二乗値(χ^2)，ｔ-score などが挙げられる（Oakes, 1998；斉藤他, 1998；山本他, 2003）。英語コーパス関連の文献にしばしば引用される自己相互情報量は意味的特性に注目した語の抽出に効果的であるとされるのに対し，ｔ-score は文法的な結合度の高い文型や機能語の抽出によいとされている（井上,1999）。広く単語検索ツールとして知られている WordSmith の機能の１つである keyness は特徴のある語彙を抽出するのに対数尤度比またはイエーツの補正公式を利用している。英語教育の分野ではまだ報告されていないようであるが，これらの複数の統計的指標を比較した研究は，自然言語処理の分野では，特定表現の抽出（久光他, 1997）や，専門用語の抽出（内山他, 2003）に応用されている。なお，本稿では，以後，論を進めるにあたって，対数尤度比などの統計的指標によってコーパスより抽出された語彙を当該英文素材の特徴のある語彙，すなわち，「特徴語」と定義する。

最近はインターネット上のデジタルデータを入手してデータベースを構築し，WordSmith 等を使用して比較的容易に単語リストを作成することができるようになった。しかし次の段階として，大量の単語リストを前に，頻度の基準だけで例えば「TOEIC に特に多く出現する語」等の当該英文素材のジャンルに特有の語を認識・選別しようとすると，多大の労力を要するという状況に直面する。そこで，大量のデータを短時間で客観的に概観できる統計的指標を有効に活用することができれば, 語彙選定の専門家でなくとも, 信頼性の高い特徴語を, 簡便かつ高精度に抽出することができることになり，各種の英文素材の特徴語リストの作成, さらには辞書の重要語の選定や段階付け等への応用が可能となる。しかしながら，統計的指標を英語教育に活用するにはそれらを適用した結果，具体的にどういう特徴語が抽出されるのか，それぞれの指標によって抽出された特徴語はお互いにどう違うのか，教師の行なう語彙選定とどう違うのかということをまず明らかにする必要がある。

2. 本研究の目的

本研究の目的は２単語間の共起関係の強さを測る８種類の統計的指標を特徴語の抽出に応用し，具体的に抽出された特徴語を観察することにより，1）各指標の相関，2）各指標間の精度の比較，3）各指標により抽出される特徴語上位の比較，について検討を試み，頻度だけでなく，対数尤度比，自己相互情報量等の統計的指標を利用したより精度の高い語彙選定の

手法の可能性を検討することである。

3. 研究の方法

本研究のおおまかな流れとして，統計的指標を利用した特徴語抽出と抽出された特徴語の検討方法をフローチャートで図 1 に示した。詳細を以下に説明する。

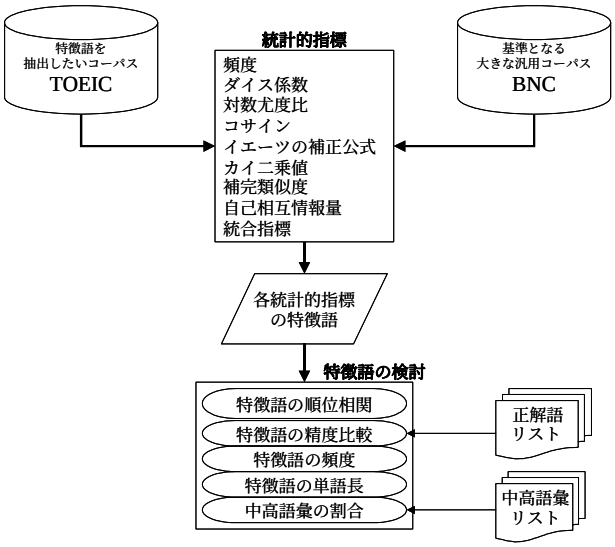

図 1　統計的指標を利用した特徴語抽出と抽出された特徴語の検討方法

3.1 言語資料

（1）　特徴語抽出資料：特徴語を抽出する言語資料には，広く日常生活とビジネスで使われる語が出現するため，その特徴語が観察しやすいといわれる TOEIC 試験模擬問題 16 回分より作成した TOEIC 語彙リスト（中條, 2003：異語数 5,016 語，延べ語数 107,077 語）を用いた。以下 TOEIC と表記する。

（2）　基準言語資料：特徴語抽出資料に出現する語の共起頻度と比較を行なう基準となる大きな汎用コーパスには British National Corpus を用い，頻度 100 以上の 38,683 語に見出し語化処理等を施した 13,995 語（Chujo, in press）を使用した。以下 BNC と表記する。

（3）　正解語リスト：実験群として統計的指標によって抽出された特徴語（候補語）の精度を評価するには，統制群として手本と想定した特徴語（正解語）との一致の度合いを調べるという手法を用いた。正解語リストには，教師が選定した「TOEIC に多く出現する語」リストである 3 レベルの学習者向け TOEIC 学習用語彙:「レベル 1(補習向け 199 語)」,「レベル 2（初級向け 200 語)」,「レベル 3（中級向け 268 語)」の計 667 語 [1)] を用いた。これらは語彙選定を専門とする英語教育者が，頻度・分布度を基準に語彙リストを作成

した後，学習の容易性・学習の必要性等の主観的判断を加え，中学校教科書出現語を除去した後，さらに，高校・大学英語教科書語彙等の外部資料との比較を繰り返して学習段階別に選択・配列したものである。学習効率も高く，信頼性の高い語彙リストであることが報告されている（中條, 2003)。本研究では中條（2003）で３段階の学習段階レベルに分類された語彙を「レベル 1」,「レベル 1+2」,「レベル 1+2+3」の 3 種類にまとめたものを使用した。

（4）中高語彙リスト：抽出された特徴語の学習段階レベルを調査するため，中学校教科書 *Horizon 1,2,3*（異語数 1,124 語，延べ語数 9,440 語），高校教科書 *Unicorn Ⅰ, Ⅱ, Reading*（異語数 3,478 語，延べ語数 36,678 語）より作成した「中高語彙リスト」を使用した。両者とも全国での採択率の高さという観点から選択した。

以上 4 種類の言語資料の語彙リストは, 屈折形を基本形に集約する見出し語化処理を施し，不偏的な特徴語の比較という目的に不要な固有名詞，数字を除去したものである。

3.2 統計的指標

本研究では，当該英文の特徴を高く反映した特徴語を抽出できると考えられる，頻度，ダイス係数（山本他, 2002），対数尤度比(池田他, 1989)，コサイン（山本他, 2002），イエーツの補正公式（池田他, 1989），カイ二乗値（池田他, 1989），補完類似度（山本他, 2002），自己相互情報量（Christopher, *et al*., 1999）の８種の統計的指標を用いた。各指標の定義を以下に述べる。表 1 に特徴語を抽出するための指標を定義する時に利用するパラメータを示す。

表 1　単語の出現状況を示すパラメータ

	TOEIC	BNC
単語X	a	b
単語X以外	c	d

a = TOEIC に単語 X が出現した回数
b = BNC に単語 X が出現した回数
c = TOEIC の延べ語数 － a
d = BNC の延べ語数 － b
N = a + b + c + d

表 1 の a,b,c,d を利用して，たとえば自己相互情報量を求めようとすれば以下の式を用いる。

$$\text{自己相互情報量} = \log\left(a\,N / \left(\left(a + b\right)\left(a + c\right)\right)\right)$$

その他の統計的指標も表 1 の a, b, c, d, N によって定義され，自己相互情報量と同様に特徴語を抽出するのに利用できるものである [2)]。

また，以上の 8 つの単独の指標（以下，単独指標）に加えて，もう１つ「統合指標」を追加した。一般的に，独立した指標の組み合せは効果的な場合が多いと報告されている（内山他, 2003)。そこで，後述の表２の４グループから１つずつ，対数尤度比，ダイス係数，自己相互情報量，コサインを選び，内山他（2003）における F_{cum} という方法によって組み合わせ

て「統合指標」とした。

3.3　指標間の相関

各単独指標の特徴語リストの順位相関によって指標をグループ分けした。

3.4 指標間の精度の比較

指標間の精度の比較は以下の方法で行なった。まず，実験群として各指標によって抽出された特徴語（候補語）を用意する。つぎに，統制群として教師が学習段階別に選定した TOEIC 学習用語彙「レベル 1」,「レベル 1+2」,「レベル 1+2+3」の 3 種類の正解語リストを利用し，候補語との一致の度合いを評価した。その評価の基準としては平均精度（内山他, 2003）を利用した。平均精度は候補語（特徴語）リストの上位から順に候補語を調べて，それが正解であったときには，そのときの（順位）精度（その順位を r としたとき，それまでの順位での正解の個数を c とすると c / r）を求めていき，リストの最後の時点で，それまでに得られた正解における（順位）精度の平均を求めたものである。平均精度は情報検索の評価などに使われており，上位に正解が多いほど大きな値となるため今回の評価の基準として適切である。

3.5　各指標ごとの特徴語上位の比較

各指標によって抽出された特徴語の上位にランクされた語を観察する。観察には，各指標による特徴語の上位に現れた語の比較，出現頻度の比較，語の長さの比較，学習段階別語彙の出現する割合を比較する方法を用いた。

4．結果と考察

4.1　指標間の相関

８種類の単独の統計的指標を用いて，各単語について，特徴語抽出資料（TOEIC）と基準言語資料（BNC）とを比較した場合での，TOEIC 出現単語としての特徴の度合を指標値として求め，その指標値で降順にソートして特徴語のリストを作成した。任意の２指標の特徴語リストについてケンドールの順位相関係数を求め，0.9 以上のものをグループ化した結果を表２に示した。

表２　順位相関による指標の分類

グループ	指　標	順位相関
A	対数尤度比, イエーツの補正公式, カイ二乗値, 補完類似度	0.9以上
B	ダイス係数, 頻度	0.9以上
C	自己相互情報量	
D	コサイン	

グループＡとＣの順位相関は 0.78～0.84，ＤとＡおよびＤとＣは 0.63～0.69，ＢとＡおよびＢとＣおよびＢとＤは 0.51 以下である。これより，対数尤度比，イエーツの補正公式，カイ二

乗値，補完類似度は互いによく似ており，ダイス係数と頻度も似ていることがわかる。また自己相互情報量とコサインはそれぞれある程度独立した指標であることがわかった。

4.2　指標間の精度の比較結果

TOEIC 用学習語彙「レベル 1」,「レベル 1+2」,「レベル 1+2+3」の３種類の語彙を正解語リストとして，８つの単独指標および統合指標がどの程度うまく正解を抽出できるかという平均精度を表３に示した。表３より統合指標の精度が最も高く，また単独指標ではダイス係数が最もすぐれていることがわかる。

表３　８つの単独指標と統合指標の平均精度

	レベル1	レベル1+2	レベル1+2+3
頻度	0.151	0.266	0.305
ダイス係数	0.168	0.289	0.321
対数尤度比	0.124	0.267	0.310
コサイン	0.068	0.151	0.220
イエーツの補正公式	0.063	0.153	0.228
カイ二乗値	0.056	0.135	0.208
補完類似度	0.055	0.134	0.205
自己相互情報量	0.042	0.096	0.157
統合指標	0.188	0.358	0.389

図２は各指標の特徴語を指標値の上位から見ていった時の正解数（縦軸）と特徴語の順位（横軸）の関係を「レベル 1+2+3」について調査した結果である。図の左側に折れ線グラフが近いほど精度が高いことを表す。また，たとえば統合指標（黒太線）を例にとると，統合指標上位 1,000 語の特徴語を検討して取捨選択すれば専門家が選んだ 426 語（正解語）（☆印）を獲得できるということがわかる。

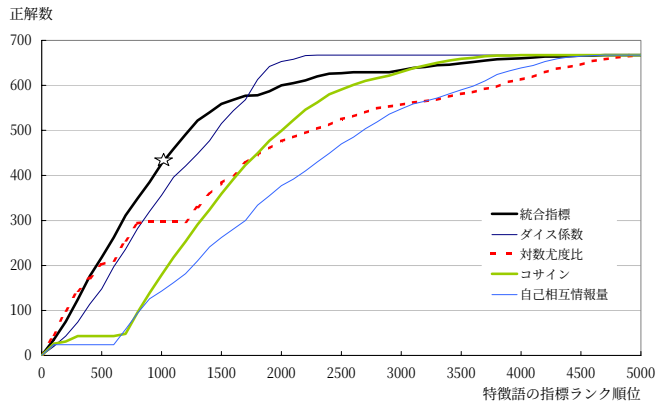

図２　特徴語の指標ランク順位と正解数

図２から指標値の上位の方では統合指標の正解数が多く，途中からダイス係数の正解数が上回る。対数尤度比については上位 800 位程度まではこれら２つの指標と同程度の正解数なので比較的少ない語数の特徴語抽出時に利用できるといえる。なお，図２に含まれない４つ

の指標については頻度はダイス係数と似た傾向であり，イエーツの補正公式，カイ二乗値，補完類似度は上位で対数尤度比がこれら３つより優れているほかは対数尤度比とほぼ同じ傾向であったので省いた。また，正解語リスト「レベル 1」,「レベル 1+2」の調査結果は図２と同様の傾向が見られたので割愛したが，表３の平均精度の観察と合わせて，統合指標とダイス係数の指標が適切な特徴語を得るために有効であることを再度確認することができた[3)]。

4.3 特徴語の比較

各指標より求められた特徴語を指標値の高いものから順に一覧表にして表４～９に示した。

表４　頻度

順位	単語	頻度
1	the	7786
2	be	5363
3	a	4115
4	to	3457
5	of	2263
6	in	1862
7	you	1604
8	will	1426
9	have	1350
10	for	1272
11	and	1197
12	I	1044
13	do	961
14	on	948
15	it	939
16	at	901
17	we	893
18	what	849
19	this	786
20	they	768
	中央値	1235

表５　ダイス係数

順位	単語	頻度
1	company	402
2	what	849
3	will	1426
4	office	312
5	question	332
6	refer	229
7	follow	318
8	new	382
9	man	349
10	you	1604
11	employee	193
12	woman	278
13	service	265
14	sale	204
15	a	4115
16	at	901
17	please	195
18	business	212
19	do	961
20	how	301
	中央値	325

表６　対数尤度比

順位	単語	頻度
1	office	312
2	refer	229
3	employee	193
4	will	1426
5	company	402
6	question	332
7	what	849
8	sale	204
9	please	195
10	hotel	155
11	customer	149
12	follow	318
13	vacation	57
14	store	129
15	computer	153
16	a	4115
17	service	265
18	business	212
19	mail	74
20	woman	278
	中央値	221

表７ イエーツの補正公式　表８ 自己相互情報量[4)]　表９　統合指標

順位	単語	頻度
1	check-out	17
2	downtown	17
3	e-mail	16
4	upcoming	15
5	hamburger	13
6	copier	11
7	ferryboat	10
8	teal	9
9	beverage	9
10	interoffice	8
11	reimburse	8
12	accordance	8
13	vacation	57
14	payload	7
15	alumni	7
16	sightseeing	7
17	salespeople	7
18	newsstand	7
19	forfeit	7
20	requisition	7
	中央値	9

順位	単語	頻度
1	cross-cultural	1
2	discontinue	2
3	cookbook	1
4	reorder	1
5	short-sleeved	2
6	comfortingly	1
7	lost-and-found	1
8	ferryboat	10
9	taxicab	1
10	carefulness	1
11	preempt	1
12	no-smoking	2
13	below-mentioned	2
14	paper-recycling	1
15	security-cleared	2
16	checkpoint	3
17	prepackaged	1
18	fabricate	2
19	conditionally	2
20	first-come-first-served	1
	中央値	1

順位	単語	頻度
1	refer	229
2	employee	193
3	office	312
4	sale	204
5	question	332
6	company	402
7	please	195
8	hotel	155
9	customer	149
10	store	129
11	computer	153
12	what	849
13	follow	318
14	will	1426
15	business	212
16	travel	125
17	service	265
18	hour	173
19	order	184
20	room	190
	中央値	200

紙幅の制限により上位 20 語のみを表示した。単語の左側にランク，右側に頻度を配した。最下段は 20 語の出現頻度の中央値である。各特徴語の指標値の表示は省略した。コサイン，イエーツの補正公式，カイ二乗値，補完類似度は上位 20 位に 1～2 語を除いて同じ語が同順

序で現れたため，イエーツの補正公式の表７に代表した。上位 20 位だけでは網羅的な比較はできないが，ある程度の傾向は観察できると考える。表４～９の最下段の出現頻度の中央値を一瞥すると表 4，5，6，7，8 の順に小さくなり，各単独指標が異なる頻度域の語を抽出していることが明らかで，各指標が異なるグループの特徴語を抽出するのに有効に働いているらしいことがわかる。まず，各指標による特徴語の上位語を指標別に見ていく。

（1）頻度（表４）

調査の規模の大小を問わず，頻度表の上位に通常現れる機能語等が並ぶ。

（2）ダイス係数（表５）

ダイス係数は頻度以外の他の指標に比べると高頻度のものが上位に来ており，特徴語を学習する前段階として基本語彙の復習が必要な補習レベル学習者に好適と思われる。

（3）対数尤度比（表６）

「数学的な背景が最も明快であり，精度もさほど悪くなく，普遍的に利用できる指標である」（久光他, 1997）といわれるが，対数尤度比は図２でも観察されたように，指標値上位ではかなり効果的に候補語（特徴語）を判別している。ダイス係数よりも TOEIC に特徴的な customer, hotel, store, vacation, computer 等の語が抽出されていることから, ダイス係数で選択される特徴語よりも一段階上のレベルで学習すると好ましい初級レベルの特徴語を抽出している。

（4）イエーツの補正公式，コサイン，カイ二乗値，補完類似度（表７）

表 7 に代表したこれら４指標には対数尤度比より少し高レベルの TOEIC 特徴語といえる interoffice, copier, upcoming, reimburse, forfeit, alumni や beverage, check-out, downtown, newsstand, hamburger 等，英語圏の文化背景と共に教えたい日常語が現れる。

（5）自己相互情報量（表８）

自己相互情報量には低頻度語彙の適用はふさわしくないといわれるように（久光他 1997），表８でも低頻度語彙が過大評価されていることがわかる。しかし，視点を変えれば，情報量の多いといわれる低頻度語は，ハイスコアを目指す上級レベルの学習者対象の特徴語と考えることもできる。

（6）統合指標（表９）

４種の単独指標（ダイス係数，対数尤度比，コサイン，自己相互情報量）を組み合わせた統合指標が抽出した特徴語には, TOEIC の特徴であるビジネス・コンテキストで広く使われる語, たとえば, 会社・人事・オフィス・出張の分野で用いられる business, company, employee, office, computer, travel, hotel, room 等の基本語，そして日常生活における sale, store, customer, service, order 等買物・購入の分野に分類される基本語がわずか 20 語にバランス良く配され，初級向けの良質な TOEIC 特徴語を抽出していると判断できる。

さらに表 4〜9 を眺めると各指標の特徴語の相違点は頻度の差だけでなく，語の長さにも表れているようであるので，次に単語の長さについて考察を進めた。表４～９で検討した指標の上位にランクされている特徴語を 50 位ごとに区切り，１位からの語の長さ（文字数）の平均を 500 位まで求めて図３に示した。結果，６種類の指標はほぼ６レベルに分かれ，自己相互情報量，イエーツの補正公式，対数尤度比，統合指標，ダイス係数，頻度の順に長い語を抽出していることが判明した。統合指標はほぼ中ほどに位置した。一般に語の長さは認知レベルの上昇とともに段階的に長くなる傾向がある（竹蓋他, 1994）ことから，各指標が異なる語彙レベルの特徴語を抽出しているらしいということが確認できた。６種の指標は，語の長さの情報を指標値の算出にまったく使用していないにもかかわらず，語彙レベルとの関連をうかがわせる結果が得られたことは興味深い。

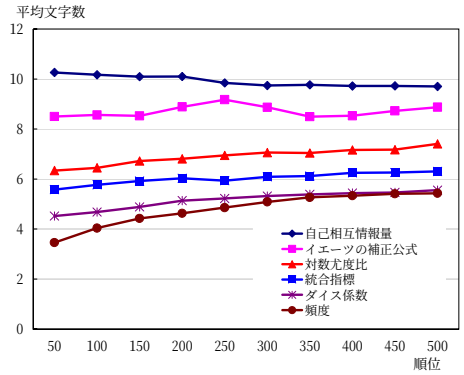

図３　各指標上位 50 位ごとの単語長の推移

そこで，もっとも各指標の特徴がよく表れていると考えられる各指標の特徴語の上位 500 語と中学校･高校教科書語彙を比較してさらに検証をすすめた。上位 500 語のうち，まず中学校教科書に出現する語，引き続き高校教科書に出現する語，そして中学・高校教科書の学習段階には出現しないそれ以上のレベルの特徴語の語数を調査し，TOEIC の主な目標語彙レベルと考えられる「中・高レベル以上の語」を中心にした結果を図４に示した。

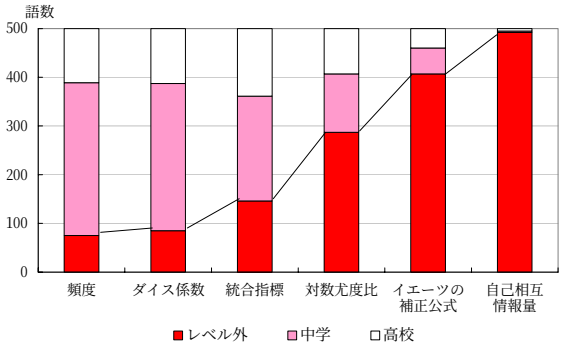

図４　中学・高校教科書に出現しない特徴語の語数

図４より，各指標は明らかに異なる学習段階レベルの語を抽出していることがわかる。指標の特徴語 500 語のうち高校３年生までの教科書語彙で未習となる（図 4 の「レベル外」の語）割合は，頻度が 15%，ダイス係数 17%，対数尤度比 57%，イエーツの補正公式が 81%，自己相互情報量が 98%，統合指標が 29%である [5)]。具体的にこの結果を TOEIC 学習用語彙の選定に応用するとすれば，ダイス係数は補習レベルの学習者向けの TOEIC 特徴語に好適であり，対数尤度比の特徴語を初級レベルとすれば，統合指標の特徴語は，補習レベルから初級レベルへの橋渡しとなる入門レベルの語彙選定にふさわしいと考えられる。イエーツの補正公式では中学･高校既習語はわずか 19%にすぎないので良質な TOEIC 中級レベルの学習語彙の特徴語を選別する。最後に，自己相互情報量の特徴語では中学・高校教科書語彙で学べる語は 500 語中わずかに 8 語だけなので，たとえば，TOEIC730 点以上を目指す上級レベル学習者向けの語彙の抽出に適しているといった具合に考えることもできよう。

以上の特徴語の比較・検討から明らかになったことは，各指標は異なる語彙レベルの特徴語を抽出していること，従って１種類の指標だけですべてのレベルの学習者のニーズを満たす語彙選択は不可能らしいこと，選定者が特徴語の使用目的・対象レベルを考慮し，これらの統計的指標を上手に使い分けて不要な語を排除すれば，頻度だけを基準として特徴語を選定するよりも少ない労力で精度の高い語彙リスト作成が可能であろうということである。

5. まとめ

調査言語資料の特性を代表する語を選定する試みは 20 世紀初頭の Thorndike の頃より欧米を中心に始まり，現代も続けられている。その際に用いられる主な基準は「よく使われる語は重要である」という理由から，頻度という客観的指標であった。しかし，頻度だけでは本当に必要な語彙，たとえば情報量の多いとされる低頻度語などが適切に選定されていない（Richards, 1970）等の指摘があり, 頻度の欠点を補うために coverage indices（Mackey & Savard, 1967），availability (Richards, 1970)，familiarity（Richards, 1974）等，主観的な基準を数量化する努力がなされてきた。しかしながら，今まで考案されたいずれの指標も「頻度」と有機的に統合させて期待される特徴語を簡便かつ高精度に抽出できる「頻度+α」の客観的基準には至っていない。

本研究はこのような状況のもと，８種の単独の統計的指標，そして８種のうち比較的独立した４指標を組み合わせた統合指標，以上の９種の統計的指標の語彙選定への応用の実用可能性を検討した。結果，以下のような点が明らかになった。

- ８種の単独指標は順位相関から見て４グループに分類される。
- 人間の行なう選定との近似という観点から９種の指標の抽出精度を比較すると，統合指標の優位性が示された。また，単独指標の中では，ダイス係数が有効であった。

- ９種の統計的指標は語彙レベル別に異なる特徴を明白に捕捉しており，学習者の語彙レベルに応じた特徴語を抽出する指標として実用的に使える可能性が極めて高い。
- もしどれか有力な指標を１つ選ぶとすれば，統合指標，あるいはダイス係数が優れていると言えるが，総合的に見て，単一の指標だけで「頻度+α」の基準には至らないであろうことも示唆された。学習者の語彙レベルや目標語彙レベルに合わせて適した指標を用いることが推奨される。

以上，本研究は統計的指標の英語教育用語彙選定への応用を試みた予備的な研究として十分に意味のある結果が得られたと考える。

謝辞　本稿をまとめるにあたり，千葉大学の西垣知佳子氏に貴重なご意見を頂きました。また，本研究の一部は（財）国際ビジネスコミュニケーション協会 TOEIC リサーチ助成制度の支援により行なわれました。ここに感謝いたします。

注

1）2）　これらは http://www5d.biglobe.ne.jp/~chujo/で公開予定である。

3）　実験で使用した３種類の正解語リストは頻度をベースに他の基準を加味して選定されたものである。頻度よりも高い精度を示した指標があったことは特筆すべきである。

4）　自己相互情報量は 630 位まで同一値だったのでランダムに順位付けた。

5）　ここでの比較は word-form による。実際には TOEIC には形は同じでも学校英語教科書とは異なる意味で用いられる語が多いので，意味も考慮すれば未習語の割合はより高くなる。さらに，用いた教科書は上級レベルの高等学校教科書であることを付言する。

参考文献

Christopher, D. Manning and H. Schutze. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Chujo, K. (in press). Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List. 『JAECS(英語コーパス学会)10 周年記念論文集』.

Mackey, W.F. & Savard, J.-G.. (1967). The Indices of Coverage: A New Dimension in Lexicometrics. *IRAL*, 2 (3) , 71-121.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.

Richards, J. C. (1970). A Psycholinguistic Measure of Vocabulary Selection. *IRAL*, 8 (2) 87-102.

Richards, J. C. (1974). Word Lists: Problem and Prospects. *RELC Journal*, 5 (2) 69-84.

池田央. (1989). 『統計ガイドブック』新曜社.

井上永幸. (1999). 「コーパスを使った語法研究と辞書編集」『英語表現研究』16, 50-60.

内山将夫・井佐原均. (2003). 「複数尺度の統計的統合法とその専門用語抽出への応用」『情報処理学会自然言語処理研究会資料』 NL157, 1-8.

齊藤俊雄・中村純作・赤野一郎. (1998). 『英語コーパス言語学－基礎と実践』研究社出版.

全国英語教育研究団体連合会. (1981). 『高校基本英単語活用集（改訂新版）』研究社出版.

園田勝英. (1996). 『大学生用英語語彙表のための基礎的研究』言語文化部研究報告叢書 7. 北海道大学言語文化部.

大学英語教育学会教材研究委員会. (1983). “JACET List of Basic Words”『「英語講読用教科書のあり方」についてのアンケート調査報告－「JACET 基本語第２次案」を中心に』18-39.

大学英語教育学会基本語改訂委員会. (2003). 『大学英語教育学会基本語リスト JACET List of 8000 Basic Words』

竹蓋幸生・中條清美. (1994). 「語彙リスト「現代英語のキーワード」－その開発と有効度の検証」『千葉大学教育学部研究紀要』42, 253-267.

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」『千葉大学英語学･言語行動研究会紀要』4, 53-63.

中條清美. (2003). 「英語初級者向け「TOEIC 語彙１，２」の選定とその効果」『日本大学生産工学部研究報告』36，27-42.

東京都中学校英語教育研究会研究部. (1986).「英語基本語彙 1,000 語, 補足 460 語, 外来語（英語）400 語」『語彙と英語教育（9）』

久光徹・丹羽芳樹. (1997). 「統計量とルールを組み合わせて有用な括弧表現を抽出する手法」『情報処理学会自然言語処理研究会資料』NL-122, 113-118.

山本英子・梅村恭司. (2002).「コーパス中の一対多関係を推定する問題における類似尺度」『自然言語処理』9-2, 45-75.

山本英子・乾裕子・井佐原均. (2003). 「主観的評価に基づく語間関係の評価尺度の比較」『言語処理学会第９回年次大会発表論文集』27-30.