論　文

日本大学生産工学部研究報告B
2006 年 6 月 第 39 巻

BNC 口語 3 分野からのレベル別 ESP 語彙の抽出

中條清美*・西垣知佳子**・内山将夫***・中村隆宏****・山﨑淳史*****

Identifying and Selecting Level-Specific ESP Vocabulary from Three BNC Spoken Components

Kiyomi CHUJO, Chikako NISHIGAKI**, Masao UTIYAMA***, Takahiro NAKAMURA**** and Atsushi YAMAZAKI******

Vocabulary development is essential to English proficiency. Chujo and Utiyama (2005, 2006) established nine statistical measures to identify domain-specific words from two 7-million-word corpora in British National Corpus (BNC) written component : commerce and applied science. This current study was undertaken to discover the value of those measures in identifying and selecting level-specific specialized vocabulary contained in the spoken component of the BNC. The study focused upon gauging the effectiveness of three 1-million-word spoken corpora for business, public/institutional, and leisure. When the nine measures were applied it was found that : (1) each statistical measure extracted a different level of domain-specific words by vocabulary level, grade level, and school textbook vocabulary coverage ; and (2) beginning-level words were identified by *cosine* and *complimentary similarity measures*, while intermediate-level words were produced by the *log-likelihood ratio*, the *chi-square test*, and the *chi-square test with Yates's correction* ; and advanced-level word lists were created with *pointwise mutual information* and *McNemar's test*. The authors show that it is helpful to use statistical tools as determiners of specialized vocabulary from 1-million-word corpora when extracting multi-level specialized lists that can be targeted to students' vocabulary proficiency levels.

キーワード：ESP，分野特徴語，統計指標，話し言葉，語彙レベル

1．はじめに

近年，学習者のニーズに符合させて目標分野を限定し，その分野特有の学習内容を指導することによって英語学習の効率を上げる ESP（English for Specific Purposes：特定目的のための英語）が注目されている。ESP は，1960 年代後半から英語が国際的な商取引や科学技術の分野で通用語となったことを背景に，本格的に発展した（Hutchinson and Waters，1987：6）[1]。ESP の定義は，「一般的な目的のための英語教育」である EGP (English for General Purposes) と区別して，学習者の「そ

*日本大学生産工学部教養・基礎科学系助教授
**千葉大学教育学部助教授
***情報通信研究機構主任研究員
****小学館
*****日本大学大学院生産工学研究科博士前期課程数理工学専攻 2 年

れぞれの学問領域や職域等で具体的目標を持って使用される英語についての研究と教育」とされている[2)~4)]。ESPの基本的特徴の1つに大量の特徴語（その分野に特有の語，分野特徴語）の存在があり（Nation，2001：204）[5)]，そのような語彙をどのように選定し，どのように指導するかがESPの課題の1つとなっている（竹蓋，1981：6；Bramki and Williams，1984：169；Baker，1988：91；Chujo and Genung，2004；Chujo and Utiyama，2006）[6)~10)]。

従来行なわれてきた語彙選定では「頻度」[11)~14)]や「レンジ」（何種類の言語資料に出現するか）の基準が用いられることが多かった[15)~21)]。頻度やレンジは主として「EGPのための語彙」，すなわち，「どの分野にも広く用いられる語彙」を選定するために有効である。しかし，これらの基準をそのままESP向けの学習語彙の選定に用いることには問題があった。なぜならば，ESP分野のテキストから作成した頻度順やレンジ順の語彙リストの上位には，頻度やレンジの特性として「どの分野にも用いられるような一般的な語彙」が多く含まれ[22)]，その中から特徴語を抽出するには，その分野に関する教材作成者の豊富な知識と経験が必要になるからである。そのため，ESPにおいては大量の分野特徴語をどのようにして精度良く簡便に選定し，効率的に学習者の語彙増強を図るかということが課題の1つとなっていた。

このような状況の中，ESP分野の特徴語選定の手法に関する先駆的研究に竹蓋（1981）がある[23)]。竹蓋はある分野Xの特徴語として「特徴語第Ⅰ類」と「特徴語第Ⅱ類」を定義した。「特徴語第Ⅰ類」とは「分野Xの語彙」から「特徴のない語彙」を除いた単語である。Xのみに出現した語彙は分野Xの明確な特徴語と考えられる。「特徴語第Ⅱ類」とは，「分野Xの語彙」と「特徴のない語彙」の両方に出現している語彙のうち，その頻度数に大きな差のある語彙を言う。特徴語第Ⅱ類の優れている点は，Ⅰ類で選定された「明確な特徴語」に続く特徴度の高い語彙を抽出できることにあり，しかも差の基準を変えることによって特徴度の強さの異なる語彙を抽出できる点にあった。このように竹蓋の研究は，特定の分野Xの特徴語を抽出するには，分野Xの語彙リストのみを見るのでは不十分で，「特徴のない語彙」を基準として比較する必要があることを示した点でそれまでの語彙研究で用いられていた手法と一線を画している[24)~26)]。

最近，「ESP分野の語彙」と「一般分野の語彙」の出現頻度を比較し，ESP分野に顕著に出現する語を比較的容易に抽出する方法として，英文検索プログラムWordSmith Toolsの*keyword*という機能を利用することができるようになった[27)]。*keyword*は対数尤度比またはイエーツの補正公式という統計指標を使っており，この*keyword*を使って特定分野のコーパスの特徴語を抽出する研究が行なわれている[28)]。例えば，Nelson（2000）は*keyword*を使って100万語のビジネス英語コーパスと200万語のBNCサンプルコーパスの出現頻度を統計的に比較し，*market*，*customer*，*management*，*price*，*bank*等のビジネス特徴語を抽出した[29)]。Flowerdew（2003）はproblem-solution patternの後に現れる特徴語を抽出し，Tribble（2000：81）は*keyword*を文体情報の抽出に利用した[30),31)]。また，対数尤度比以外にも，自己相互情報量やt-scoreなどの統計指標を利用して各単語に独自の値を与えて特徴度を示す研究が増加している（Oakes，1998）[32),33),34)]。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指標によってコーパスより抽出された語彙をある特定分野の「特徴語」と定義する。

このようにコーパス言語学において特徴語を抽出するための特徴度を示す様々な統計指標が考案され，利用されるようになってきた。しかしながら，それらの統計指標が具体的にどのような単語に高い指標値を与えているのか，ある統計指標で得られる指標値と他の統計指標で得られる指標値が互いにどの程度似ているのか，似ていないのか，異なる統計指標がどのような点でどのように「良い」指標であるかということは，これまでほとんど明らかにされていなかった。

これに対し，中條・内山（2004），内山・中條・山本・井佐原（2004）では9種類の統計指標（頻度，ダイス係数，対数尤度比，コサイン，イエーツの補正公式，カイ二乗値，補完類似度，自己相互情報量，統合指標）を10万語からなる1種類の言語資料の特徴語抽出に適用し，指標どうしの類似度や各指標の抽出精度の検証（語彙選定を専門とする英語教育者の選定した正解語リストとの比較）を行なった[35),36)]。その結果，各指標は有効な精度で，初級・中級・上級といった異なる語彙レベルの特徴語を抽出している可能性が確認された。

中條・内山・長谷川（2005）では，教育現場での実践的利用の試みとして，それぞれ2万語からなる小規模な時事英語資料18種（TIME，CNN他）の特徴語抽出に統計指標を適用し，上位に順位付けられた特徴語の比較，検討を行なった。結果，各統計指標は，2万語という比較的小規模の資料においても，また1編のニュースにおいても，それぞれ初級，中級，上級の語彙レベル別に特徴語を抽出していることが明らかになった[37)]。

続いて，Chujo and Utiyama（2005）とChujo and Utiyama（2006）では規模を拡大して，BNCの書き言葉の部（written component）から，Applied ScienceとCommerce分野の各700万語のコーパスに統計指標を適用し，これらの指標が比較的大規模な書き言葉の言語資料からの語彙レベル別ESP語彙抽出に有効であることを示した[38),39)]。

本稿は，これまでに行なってきたESP語彙選定方法開発プロジェクトの一環として，今日の英語教育においてその指導が強く望まれている話し言葉におけるESP語彙の抽出を試みた。使用した言語材料はBNCの話し言葉の部(spoken component)のBusiness, Public/Institutional, Leisureというそれぞれ100万語超からなる3種類のESPコーパスである。3分野のESP語彙抽出に9種の統計指標を適用することによって,「100万語規模」の「話し言葉」コーパスの資料においても初級，中級，上級の語彙レベル別のESP語彙を有効に抽出できるかを検証するものである。

2．研究方法

本研究の研究手順の概要をフローチャートに示した(**図1**)。研究の手順は，1) 統計指標を利用した特徴語抽出，2) 抽出された分野特徴語の検討，の2部から構成される。詳細を以下に説明する。

2.1 特徴語抽出に使用した言語資料[注1)]

(1) ESPコーパス（特徴語を抽出したい資料）

特徴語を抽出するために使用したESPコーパスは，

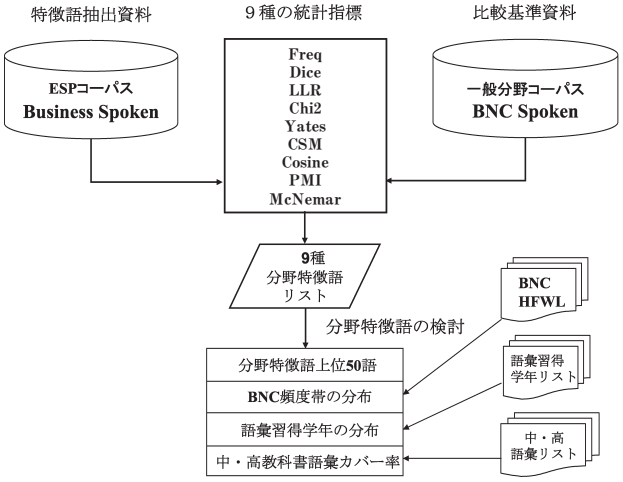

図1 統計指標を利用した特徴語抽出と抽出された特徴語の検討方法

BNCの「話し言葉の部」を構成するBusiness, Public/Institutional, Leisureの3分野である（**表1**）。3種のESPコーパスは**表1**中の「カテゴリー」欄に示したコンテクストにおいて収集された発話で構成され，各132万語，134万語，146万語のテキストからなる[注2)]。最初に，3種のESPコーパスから，3種の分野別の語彙リストを作成した。これらは，CLAWSによるタグ付けとレマタイズ・プログラムを利用して単語の屈折形とその頻度を基本形に集約したレマリストである[40),41)]。次に，レマリストから頻度10以上の語彙を抽出し，固有名詞と数字等を人手で除外した。**表1**には固有名詞と数字を除外した後の頻度10以上のレマ数を記した。

(2) 一般分野のコーパス（基準となる大きな汎用の言語資料）

特徴語の抽出においては，ESPコーパスに出現する語の生起頻度と，その語の一般分野のコーパス資料での生起頻度とを比較し，その出現の度合が大きく異なるものを特徴語の候補と考える。そのような比較のための基準となる一般分野のコーパスには，BNC話し言葉1,036万語を使用し，そこから求めた頻度10以上の8,462語（延べ語数9,126,606語）を以下BNC Spokenリストと表記して用いた。ただし，この8,462語には上述の3種のESPリストが一部に含まれているので，特徴語抽出の際には該当する各ESPリストを除いたBNC Spokenリストを用いた[注3)]。

2.2 特徴語の検討に使用した言語資料

抽出された分野特徴語の語彙レベルがどのようなものかを観察するために，以下の3種の資料を用いた。

(1) BNC High Frequency Word List (BNC HFWL)

抽出された語の頻度分布を評価するため，BNC HFWL (Chujo, 2004) を使用した[42)]。BNC HFWLは，BNCの書き言葉，話し言葉のすべてを含めた全語彙のうち，固有名詞と数字を除外した，頻度100以上に該当する13,994語（延べ語数86,112,272語）のレマリストである。

(2) 英語母語話者の語彙習得学年リスト

表1 調査したBNCの話し言葉3分野の言語資料

ESP分野	延べ語数	頻度10以上のレマ数	カテゴリー
Business	132万語	2,780語	company talks and interviews, trade union talks, sales demonstrations meetings, consultations
Public/ Institutional	134万語	3,331語	political speeches, sermons, public/government talks, council meetings, religious meetings, parliamentary proceedings, legal proceedings
Leisure	146万語	3,229語	speeches, sports commentaries, talks to clubs, broadcast chat shows and phone-ins, club meetings

抽出された語の学年分布を見るため，Dale and O'Rourke（1981）の *The Living Word Vocabulary* を参照した[43]。この資料は 40,400 項目の語の意味を，米国人の 75%以上の子供が理解できる学年を調査したものである。4 ～16 年生の学年配当の決定にはこの資料を参照した。しかし，1 ～ 3 年生の学年は区分されていないので，これらの学年の決定には Harris and Jacobson（1972）の *Basic Elementary Reading Vocabularies* の学年配当リストを参照した[44]。

⑶　中学校・高等学校教科書語彙リスト

我が国の学校英語教科書から見た語彙レベルを調査するため，中学校教科書 *New Horizon English Course 1, 2, 3*（笠島他，2002）と高等学校教科書 *Unicorn English Course I*，*II*，*Unicorn English Reading*（市川他，2003）より作成した「中・高語彙リスト」（異語数 3,245 語，延べ語数 38,937 語）を使用した[45),46)]。これらの教科書の選定に際しては，できるだけ一般的な傾向が得られるように，全国での採択率の高いものから選択した[47),48)]。

2.3　統計指標

使用した統計指標は，頻度 (Freq : frequency)，ダイス係数 (Dice : Dice coefficient)[49]，コサイン (cosine)[50]，補完類似度 (CSM : complementary similarity measure)[51]，対数尤度比 (LLR : log-likelihood ratio)[52]，カイ二乗値 (Chi2 : chi-square test)[53]，イエーツの補正公式 (Yates : chi-square test with Yates's correction)[54]，自己相互情報量 (PMI : pointwise mutual information)[55]，マクネマー (McNemar : McNemar's test)[56]の 9 種である。これらの指標は**表 2** に示したパラメタ a, b, c，d によって計算される。

このようなパラメタを利用して計算を行なう背景には，もし単語 α の ESP 分野の話し言葉における出現状況が，一般分野の話し言葉である BNC Spoken における出現状況よりも顕著であるならば，その単語 α は ESP 話し言葉分野において特徴的な単語であろうという期待がある。各指標はそのような「顕著性の度合」を測定するために利用されている。**表 2** の a，b，c，d を用いて，たとえば自己相互情報量の指標値は

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

で求められる。その他の指標も**表 2** のパラメタを用いて **Appendix** に示した定義式によって求められる。定義式における a，b，c，d の使い方は 9 種の統計指標それぞれによって異なるため，同一の単語 α であっても異なる指標値が与えられる。各統計指標の工学的な説明を本稿に加えることが望ましいが，紙幅の関係で不可能であるため，内山他（2004）(http://www2.nict.go.jp/jt/a132/members/mutiyama/pdf/chara.pdf) を参照されたい[57]。

表 2　単語の出現状況を示すパラメタ

	ESP リスト	BNC Spoken リスト
単語 α	a	b
単語 α 以外	c	d

a＝ESP リストに単語 α が出現した回数
b＝BNC Spoken リストに単語 α が出現した回数
c＝ESP リストの延べ語数－a
d＝BNC Spoken リストの延べ語数－b
n＝a＋b＋c＋d

2.4　特徴語リストの作成

9 種の統計指標を用いて，3 種 ESP リスト中の各語の出現状況を，BNC Spoken リストでの出現状況と比較した指標値を求めた。その値に従って降順に単語をソートし，特徴語リスト（9 指標× 3 資料）を作成した。

2.5　各指標の上位に順位付けられた特徴語の比較

各指標の上位に順位付けられた特徴語を評価するために，1）抽出された最上位 50 語の傾向，2）抽出された最上位 500 語についての BNC 頻度，学年分布，学習語彙から見た語彙レベルの検討を行なった。

3．結　果

3.1　上位に順位付けられた特徴語の傾向

表 3 に BNC の Business 分野，**表 4** に BNC の Public/Institutional 分野，**表 5** に BNC の Leisure 分野の特徴語を示した。紙幅の関係で，9 種の指標より求められた分野特徴語のうち指標値の高いものから上位 50 語を示した。上位 50 語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。各特徴語の指標値の表示は省略した。表の下 2 段には 50 語の出現度数の平均値と平均文字数を示した。なお，Freq/Dice，Chi2/Yates は上位 50 位に同じ語が現れたため同じ列に示した。また，特徴語の指標値が同一値の場合は ABC 順に順位付けた[注4]。

表中下から 2 段目の「平均出現度数」は，**表 3**，**4**，**5** とも同じ傾向を示し，指標が左から右に移るにつれて，出現度数が段階的に小さくなっている。つまり，Freq/Dice, Cosine, CSM, LLR, Chi2/Yates, PMI, McNemar の順に，**表 3** (Business) では 14139 → 13899 → 5908 → 3474 → 2992 → 23 → 11 と小さくなっており，**表 4** (Public/Institutional)，**表 5** (Leisure) も同様の傾向を示している。

一方，表中最下段に示した単語の文字数で見た「平均単語長」は，指標が左から右に移るにつれて，ほぼ段階的に長くなっている。例えば，**表 3** (Business) では 3.2 → 3.5 → 4.9 → 6.7 → 6.9 → 8.3 → 8.1 と長くなっており，**表 4** (Public/Institutional)，**表 5** (Leisure) も同様の傾向を示している。一般に語の長さは認知レベルの上昇

表 3　BNC Business Spoken Component の分野特徴語上位 50 語

	Freq, Dice	Cosine	CSM	LLR	Chi2, Yates	PMI	McNemar
1	be	be	we	we	we	abolition	accrue
2	the	the	the	motion	motion	accrue	arena
3	I	we	to	region	congress	acquisition	carpenter
4	you	to	that	congress	region	activist	cost-effective
5	to	that	of	union	union	adapt	craftsman
6	and	you	right	colleague	colleague	adoption	feasibility
7	that	and	if	client	client	affiliation	handshake
8	it	I	need	company	company	amalgamation	invaluable
9	we	a	okay	train	conference	arena	invasive
10	a	of	for	conference	train	arousal	marquis
11	have	have	actually	procedure	procedure	aspiration	refund
12	of	it	which	business	premium	assertive	sailor
13	do	do	because	premium	file	assertiveness	suitably
14	not	in	this	file	trade	audio	transplant
15	in	will	year	trade	business	banker	unprotected
16	they	not	work	need	composite	banner	upturn
17	will	they	will	to	delegate	batch	batch
18	get	this	company	okay	sail	beginner	beginner
19	this	for	train	quality	quality	byte	catchment
20	on	on	union	composite	project	carpenter	civilize
21	for	get	motion	project	need	catchment	fieldwork
22	what	can	thing	the	product	cedar	handbook
23	can	if	business	delegate	okay	certification	hunter
24	but	what	as	sail	spreadsheet	civilize	introductory
25	go	but	region	product	safety	clarity	marginally
26	so	right	may	right	profit	competence	organizer
27	there	so	other	member	to	composite	retailer
28	if	with	problem	safety	president	comrade	right-hand
29	say	go	member	profit	organization	concession	stressful
30	know	there	client	organization	member	conductivity	supportive
31	with	at	with	spreadsheet	appointment	consignment	yearly
32	he	think	colleague	president	employer	cost-effective	adapt
33	think	about	or	actually	assignment	craftsman	adoption
34	right	or	trade	appointment	presentation	cubic	diligence
35	at	know	job	of	right	dagger	empowerment
36	well	say	congress	manager	worker	dependant	extract
37	about	as	so	worker	branch	dignity	install
38	all	which	about	employer	the	diligence	scandal
39	or	all	month	branch	manager	dismissal	submissive
40	just	because	also	benefit	feedback	earner	whit
41	as	need	conference	presentation	actually	edit	worksheet
42	then	okay	pay	meeting	benefit	empowerment	affiliation
43	which	thing	ask	assignment	activist	entitlement	aspiration
44	now	motion	point	team	objective	excellence	cedar
45	thing	just	cost	problem	formally	extract	clarity
46	because	well	procedure	copy	advertise	feasibility	competence
47	mean	region	question	month	meeting	fieldwork	edit
48	like	actually	project	objective	manual	forefront	excellence
49	very	year	quality	management	copy	gala	gala
50	up	work	form	section	team	graduate	tabloid
平均出現度数	14139	13899	5908	3474	2992	23	11
平均単語長	3.2	3.5	4.9	6.7	6.9	8.3	8.1

表 4　BNC Public/Institutional Spoken Component の分野特徴語上位 50 語

	Freq, Dice	Cosine	CSM	LLR	Chi2, Yates	PMI	McNemar
1	be	the	the	the	the	accessibility	surrounding
2	the	be	of	council	council	accommodate	agility
3	to	to	to	of	county	accountancy	await
4	that	that	that	county	of	accuracy	believer
5	I	of	we	chairman	chairman	acknowledgement	bureaucratic
6	and	and	in	plan	plan	actuarial	causeway
7	of	I	which	settlement	settlement	actuary	centrally
8	a	we	as	lord	lord	adjoin	coalfield
9	it	a	this	committee	committee	adjourn	collectively
10	you	in	council	district	district	aerospace	condemnation
11	have	have	for	policy	policy	agility	continuation
12	we	it	county	which	authority	amenity	dedicate
13	in	you	by	authority	development	amnesty	dense
14	not	will	will	development	which	annuity	doubtful
15	will	this	plan	to	local	anticipate	drab
16	do	for	point	local	figure	apostle	elector
17	they	not	thank	that	greenbelt	applicable	eventual
18	this	council	policy	we	amendment	applicant	faithful
19	on	on	area	figure	fund	arbitrary	favorable
20	for	they	from	greenbelt	to	asylum	fullness
21	can	which	local	as	proposal	authorization	gradual
22	there	as	committee	amendment	that	await	helper
23	but	do	and	fund	service	backlog	humanity
24	he	can	make	point	budget	baptism	imminent
25	think	county	chairman	service	member	baptize	inaccurate
26	what	there	service	member	we	bedtime	incidence
27	if	if	shall	area	area	believer	indigenous
28	get	but	member	proposal	site	beneficiary	invisible
29	with	with	authority	budget	point	bridal	kindly
30	as	think	district	site	officer	budgetary	kindness
31	say	plan	actually	officer	propose	built-up	maturity
32	go	at	figure	issue	as	bureaucracy	mischief
33	at	by	government	propose	item	bureaucratic	needless
34	which	chairman	development	item	issue	burglary	omit
35	so	from	question	thank	refer	capability	onus
36	or	or	lord	government	paragraph	causeway	optional
37	know	committee	settlement	refer	provision	cease	outward
38	about	policy	need	within	plaintiff	centenary	periphery
39	all	point	also	agree	criterion	centrally	physiotherapy
40	well	settlement	report	plaintiff	government	charitable	preach
41	from	make	fact	paragraph	scheme	civic	preamble
42	just	say	within	provision	within	coalfield	preclude
43	very	lord	far	scheme	agree	collectively	printout
44	now	district	issue	report	thank	commence	psalm
45	by	what	agree	criterion	boundary	commencement	radial
46	make	very	may	concern	concern	commissioner	rash
47	right	local	new	comment	comment	committal	reliance
48	because	authority	officer	in	report	compliance	relocation
49	come	thank	year	boundary	hectare	comprise	reputable
50	then	about	fund	by	panel	conceal	testimony
平均出現度数	17760	16802	9033	6812	6209	23	10
平均単語長	3.2	4.0	5.1	6.1	6.2	8.7	8.0

表 5　BNC Leisure Spoken Component の分野特徴語上位 50 語

	Freq, Dice	Cosine	CSM	LLR	Chi2, Yates	PMI	McNemar
1	be	the	and	and	and	aboard	marking
2	the	be	the	use	use	aisle	ale
3	and	and	they	the	the	ale	anger
4	I	a	a	they	woman	ammunition	apiece
5	you	to	in	woman	they	anchor	arrival
6	to	I	use	man	nil	anger	ballast
7	a	you	to	pound	man	anorexia	brew
8	that	they	at	nil	pound	apiece	bye
9	it	that	there	at	flat	apprenticeship	concerto
10	have	have	of	lot	war	archway	dial
11	they	in	for	flat	lot	arrival	furious
12	of	of	pound	war	goal	artillery	furrow
13	in	it	from	street	street	ashore	graveyard
14	we	we	lot	a	bid	athletic	groom
15	do	do	man	goal	at	ballast	hostel
16	not	there	very	father	unite	barbed	loyalty
17	he	for	when	live	father	batsman	marketplace
18	will	he	see	bid	live	bidder	nationality
19	there	not	woman	unite	advance	blacksmith	navigator
20	on	use	as	from	slate	bleak	organist
21	for	at	work	slate	pit	bonfire	outskirts
22	get	on	day	advance	shilling	boost	pavilion
23	but	but	with	in	quarry	bowler	plank
24	go	will	then	after	coal	bracelet	prosperity
25	this	know	but	pit	a	brew	referendum
26	know	go	call	coal	yard	brilliantly	repository
27	at	this	after	shilling	from	bronze	sewer
28	can	with	about	quarry	after	bye	stump
29	what	get	any	call	mother	cabin	trumpet
30	think	think	old	old	family	calf	vandalism
31	with	then	live	there	call	caller	walkway
32	so	see	who	yard	in	cannabis	worldwide
33	well	about	many	family	old	cargo	anchor
34	say	all	child	mother	strike	churchyard	anorexia
35	all	from	or	day	boat	civilian	batsman
36	then	so	year	child	dock	classify	brilliantly
37	about	as	war	young	young	classless	classify
38	see	when	job	shop	day	coffin	compensate
39	use	well	come	boat	there	colliery	curate
40	or	or	time	job	shop	compensate	dismiss
41	as	can	we	strike	child	concerto	disturbance
42	if	very	house	ball	reserve	contraception	dyke
43	when	come	flat	horse	ball	crane	exposure
44	come	what	school	work	drug	curate	fleet
45	just	pound	into	dock	horse	daytime	footballer
46	like	say	all	drug	job	defender	greyhound
47	very	lot	shop	reserve	miner	desktop	hearse
48	from	man	much	very	guild	dial	hormone
49	up	now	father	for	server	diameter	inherent
50	now	up	street	game	coffin	disciplinary	innovation
平均出現度数	17163	17013	9248	5932	5541	22	10
平均単語長	3.0	3.1	3.6	4.2	4.4	7.1	7.4

とともに段階的に長くなる傾向があることから[58]，指標が右にいくに従い，難易度の高い特徴語が抽出されていることが推測できる。

表3，**表4**，**表5**に示した3種の分野における「分野特徴語」の上位50語は，同一のデータから抽出されたものであるが，使用した指標によって上位に順位付けられた特徴語はかなり異なっている。以下ではBusiness分野について9種の指標別に得られた特徴語上位50語を観察する。

⑴　頻度，ダイス係数(Freq/Dice)，コサイン(Cosine)

これらの指標で抽出された高頻度語(*be*，*the*，*I*，*you*，*to*，*and*，*that*，*it*，*we*，*a*，*have*，*of*，*do*，*not*，*in*，*they*，*will*，*get*，*this*，*on*，*for*，*what*，*can*，*but*，*go*，*so*，*there*，*if*)は，調査の規模を問わず頻度リストの上位に通常現れる機能語である。リストの上位には機能語に続いて，*right*，*think*，*know*，*because*，*okay*，*just*，*well*，*actually*等の内容語が並ぶ。これらの語は通常の書き言葉の頻度リストよりも上位に現れている[注5]。よく観察すると，実際にはこれらは内容語というよりも話し言葉に特有な*you know*，*I think*などのinterpersonal phrasesや*well*，*right*などのsingle-word organizational markers(Schmitt，2000：73)であることがわかる[59]。50位の終わりぐらいからmotion, regionなど初級レベルのビジネス関連用語が現れる。

⑵　補完類似度(CSM)

CSMでは上位20位くらいまでは機能語が大半を占めるが，20位あたりから*company*，*union*，*business*，*problem*，*client*，*colleague*，*trade*等の初級レベルのビジネス語が抽出されている。

他の話し言葉分野と比較して*we*がBusiness分野の特徴語として現れているのは非常に興味深い。BNCのBusiness分野コーパスには労働組合のスピーチや管理職の会議等のテキストが多く含まれるため*we*が多用されていることがわかる。Handford(2005)は100万語のCANBECビジネスコーパスの分析において[注6]，特に*we*の用法に注目した[注7),60]。Handford(2005)によると*we*は特定の「私たち」の中に聞き手を含めたり，時に除外することによって聞き手との距離感や力関係を効果的に表現したり，「同朋意識を喚起」するために効果的に多用されるという[61]。なお，*we*はLLRとChi2/Yatesでも1位に抽出されている。

⑶　対数尤度比(LLR)，カイ二乗値，イエーツの補正公式(Chi2/Yates)

これらの指標は分野特徴語の上位に，初級～中級向けのビジネス用語を精度良く抽出している。例えば，*motion*，*region*，*congress*，*union*，*colleague*，*client*，*company*，*train*，*conference*，*procedure*，*business*を始め，上位50語のほぼ半数は初級から中級のビジネス用語である。

⑷　自己相互情報量(PMI)，マクネマー(McNemar)

これらの指標は上級者向けのビジネス用語を抽出している。例えば上位には*accrue*(発生する)，*abolition*(廃止)，*acquisition*(買収)，*cost-effective*(費用効果がある)，*adapt*(適合)，*adoption*(採用)，*feasibility*(採算性)，*affiliation*(提携)，*handshake*(握手，退職金)，*amalgamation*(合併)等の専門的な語彙が並んでいる。

以上の考察は主観的な観察であるので，次節では，客観的にどのような語彙レベルの語が上位に特徴語として抽出されているかを把握するため，3種の参考資料を利用して各指標上位500語の語彙レベルを検討した。

3.2　BNC高頻度語の分布

9種の各指標が，一般分野の英語コーパスのどのよう

表6　3分野の特徴語上位500語のBNC頻度帯の分布

BNC頻度帯	Business							Public/Institutional							Leisure						
	Freq Dice	Cosine	CSM	LLR	Chi2 Yates	PMI	McNemar	Freq Dice	Cosine	CSM	LLR	Chi2 Yates	PMI	McNemar	Freq Dice	Cosine	CSM	LLR	Chi2 Yates	PMI	McNemar
1000	93.2	80.6	70.2	42.6	37.8	9.4	0.2	92.4	79.4	68.6	49.6	46.2	1.4		92.2	77.4	59.0	31.2	27.2	2.6	
2000	6.0	10.0	19.8	18.0	17.2	14.0	5.4	6.0	11.6	20.6	22.4	22.6	7.4	2.8	7.2	11.6	21.4	15.0	14.4	8.0	2.2
3000	0.8	3.8	4.8	9.6	10.4	17.8	20.2	1.2	4.6	6.2	10.8	11.2	12.4	10.2	0.4	3.4	7.6	11.2	12.4	16.6	14.6
4000		1.8	2.2	6.2	7.2	15.0	20.0		2.4	2.6	6.8	7.2	17.6	19.6		2.2	4.4	10.0	10.2	17.0	21.8
5000		1.2	1.2	7.2	7.6	15.8	20.6	0.2	0.8	1.0	3.2	4.0	15.2	19.0		1.8	3.0	9.6	10.6	15.6	19.0
6000		1.2	0.8	4.8	5.6	7.4	9.6		0.6	0.4	2.6	3.4	14.4	15.0	0.2	1.4	2.0	7.2	7.6	11.6	11.4
7000		0.2	0.2	3.8	4.2	5.8	8.0	0.2	0.4	0.4	1.6	2.0	11.2	12.0		0.8	1.0	5.0	5.2	9.4	10.6
8000				2.0	2.6	3.4	3.8				1.4	1.4	6.6	6.6		0.8	0.8	4.2	4.6	7.0	7.4
9000		0.2	0.2	1.6	2.4	3.6	4.0				0.6	0.8	3.8	4.4		0.2	0.2	1.6	2.2	4.0	3.8
10000		0.2		1.2	1.2	2.2	2.8					0.2	2.6	2.8		0.2	0.2	1.4	1.6	2.2	2.8
11000		0.2		0.4	0.4	1.2	1.2		0.2	0.2	0.4	0.4	3.2	3.0				1.0	1.2	1.6	1.6
12000		0.6	0.6	1.6	2.0	2.6	2.4				0.4	0.4	2.0	2.2				0.8	0.8	1.4	1.4
13000				0.8	0.8	0.8	0.8				0.2	0.2	1.0	1.0		0.2	0.4	1.2	1.2	1.8	2.0
13994				0.2	0.6	1.0	1.0						1.2	1.4				0.6	0.8	1.2	1.4

＞5.0%

な頻度帯に属する単語を抽出しているかを調査した。BNC HFWL を頻度順に上位から 1,000 語ずつ 14 段階の頻度帯に区切り，3 種の話し言葉の ESP 分野特徴語 500 語の各語が BNC の何千語レベルの頻度帯に属するかの割合を示した（**表 6**）。傾向を把握しやすいように各指標の特徴語の 5 %以上が属する頻度帯をグレーで示した。また各指標の特徴語が相対的に多く属する頻度帯の%の値を□で囲った。

3 分野を通じて各指標の特徴語が多く属する頻度帯を見ていく。Freq/Dice によって抽出された特徴語は 90%超が BNC 1,000 語レベルに属している。Cosine も 80%前後が BNC 1,000 語レベルに属している。CSM の特徴語の中心はやはり BNC 1,000 語にあるが，約 20%は BNC 2,000 語レベルに，5 %前後が 3,000 語レベルに属する。LLR/Chi2/Yates では抽出範囲が広がり，抽出の中心は Business では 6,000 語まで，Public/Institutional では 4,000 語まで，Leisure では 7,000 語までに属している。以上の指標では，抽出の範囲は広がっていてもその抽出の最多頻度帯は依然 BNC 1,000 語レベルであった。しかし，PMI と McNemar の特徴語の中心は，Business と Leisure では BNC 3,000～5,000 語レベルであり，Public/Institutional では BNC 4,000～6,000 語にある。PMI と McNemar のように上級レベルの語彙を抽出できる指標は今後活用が期待できる。なぜならば，語彙指導の分野では Nation が唱導する 2,000 語が一応の語彙学習の目標 (benchmark) とみなされているようであるが，それ以降の上級レベルについては「何語を目標にすればよいのか」「どのようにして選定すればよいのか」等についてまだほとんど解明されていないからである[62),63)]。

3.3 特徴語の学年分布

上位 500 語の特徴語を容易に理解できるのは，米国人の場合，どの学年ぐらいなのかを Harris & Jacobson (1972) と Dale & O'Rourke (1981) の資料に基づいて調査した。**図 2**，**図 3**，**図 4** は 3 分野の特徴語 500 語のうち何%が 1 年～16 年の各学年で理解されるかを示している。資料に収集されていない語の割合は N/A とした。仮に理解される特徴語が 8 割に達する学年を表中に

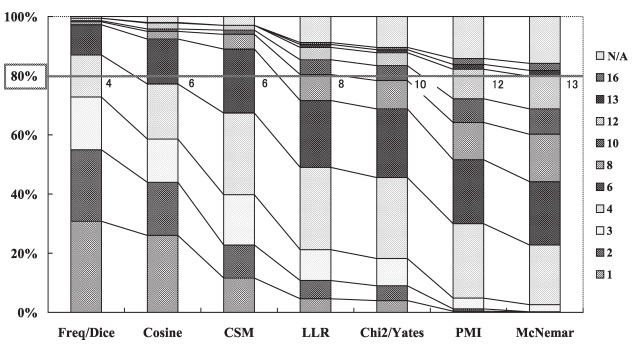

図 2　語彙習得学年分布 (Business)

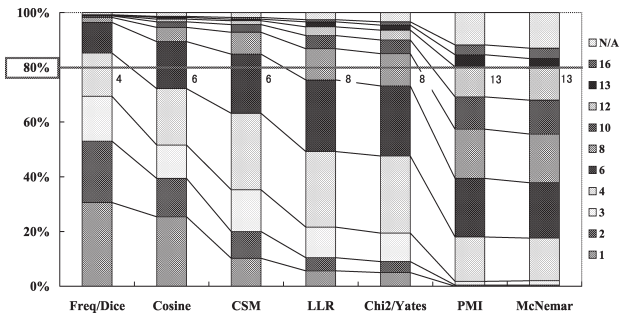

図 3　語彙習得学年分布 (Public/Institutional)

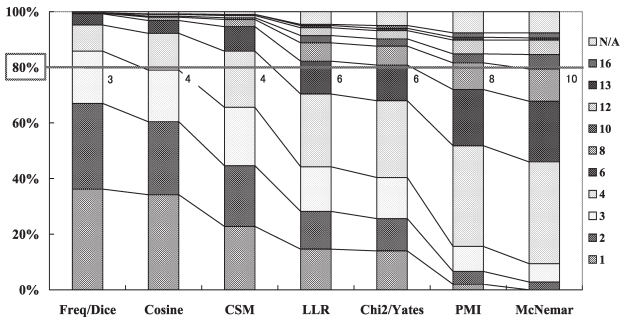

図 4　語彙習得学年分布 (Leisure)

示した。これらの学年を基準にして見ると，**図 2** の Business 分野では Freq/Dice によって抽出された特徴語は 4 年生，Cosine と CSM は 6 年生，LLR は 8 年生，Chi2/Yates は 10 年生，PMI は 12 年生，McNemar では 13 年生となる。**図 3** の Public/Institutional 分野では Freq/Dice によって抽出された特徴語は 4 年生，Cosine と CSM は 6 年生，LLR と Chi2/Yates は 8 年生，PMI と McNemar では 13 年生となる。**図 4** の Leisure 分野では Freq/Dice によって抽出された特徴語は 3 年生，Cosine と CSM は 4 年生，LLR と Chi2/Yates は 6 年生，PMI は 8 年生，McNemar では 10 年生となる。参考資料に含まれていない特徴語があるものの，各特徴語について母語話者が理解できるようになる学年の目安が得られたことから，9 種の統計指標は学習者の習熟度別学習用特徴語抽出に利用できると考えられる。

3.4 学校英語教科書語彙との差

ESP 分野の言語資料を教材として使用する際には，一般的には大学生が対象となる。そこで大学入学時の学習者の語彙レベルと ESP 語彙の関連を検討する必要がある。大学入学時の学習者の語彙レベルを，高校 3 年生までに学習する中高教科書語彙 (JSH) のレベルと仮定して，**図 5**，**図 6**，**図 7** は 3 分野の特徴語 500 語のうち何%の語が JSH で既習となるかの割合を示している。3 分野を総合して見ると，おおよそ Freq/Dice の抽出結果の 90～96%，Cosine 77～88%，CSM 71～77%，LLR 48～55%，Chi 2/Yates 44～52%，PMI 14～26%，McNemar 11～19%が既習と考えられる[註8)]。この結果からも，各指標は明白に異なる習熟度の特徴語を抽出して

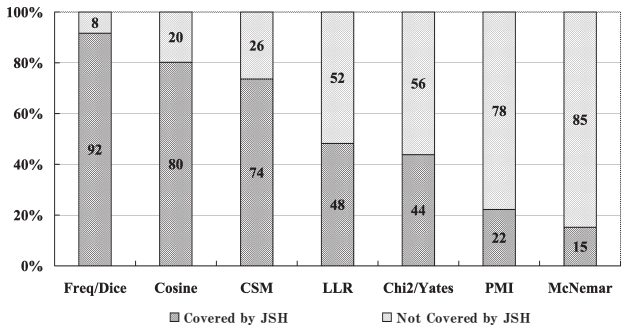

図5　中・高教科書語彙の比率 (Business)

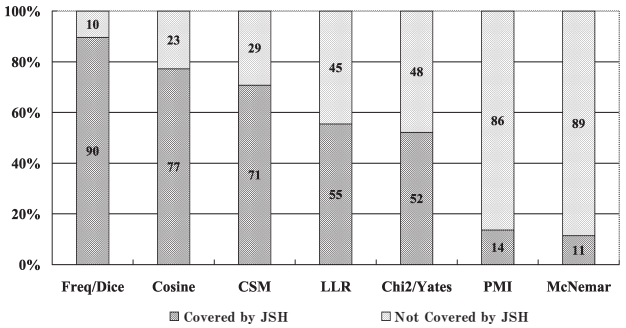

図6　中・高教科書語彙の比率 (Public/Institutional)

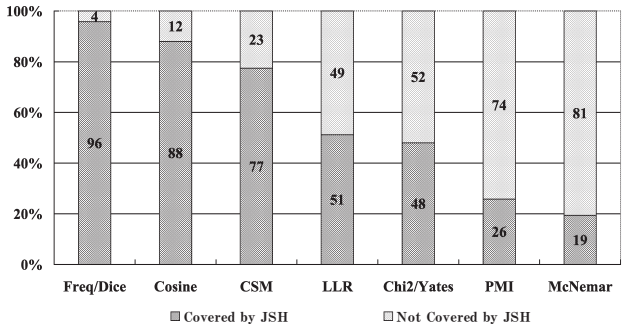

図7　中・高教科書語彙の比率 (Leisure)

いることがわかる。

4．教育用語彙選定への応用

本稿で目指している統計指標を利用した語彙選定の位置付けについて，我々は次の2段階の方法を想定している。まず，1）ESP分野の単語をその特徴度によって順位付ける，2）順位付けられた単語の上位から，選定者が教育的配慮などの主観や経験に基づいて重要な単語を選定する。本研究は2）の主観的な語彙選定の段階を効率的に実施するために，1）の順位付けの資料を精度良く機械的に得ることを目指すものである。

この前提に立って，3.1以降の考察から明らかになった情報を総合して，各統計指標を以下に述べる英語習熟度に応じた学習用語彙選定に利用できるのではないかと考えている。

①　頻度とダイス係数による抽出結果は約9割が既に学習した語であり，しかも大部分がBNC 1,000語レベルに属する語である。従って，これらの指標による特徴語は，ESP分野の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「レメディアル（補習）レベル」の学習者に適していると考えられる。

②　コサインと補完類似度による特徴語はBNC 2,000語レベルまでの語が中心であり，特徴語の約8割が，母語話者の6年生が理解するような比較的易しいと考えられる語が多い。3割近くある中学・高校教科書語彙の未習語も，BNC 3,000語レベルまでに属する基本的な分野特徴語とみなされるので，総合すると，「初級レベル」学習者向けの特徴語抽出の指標に利用できると考えられる。

③　対数尤度比，カイ二乗値，イエーツの補正公式ではBNC 7,000～8,000語レベルにまで特徴語が分布している。特徴語の約8割を母語話者の10年生で理解することができる。さらに，中学・高校教科書語彙の既習語が5割前後を占めるが，新しく学習しなければならない語が5割を占めるため，学習負荷が大きいことから，「中級レベル」学習者対象の特徴語の抽出に適していると考えられる。

④　自己相互情報量とマクネマーで抽出される特徴語の9割近い語は中学・高校教科書に出現しない語であること，母語話者の12，13年生でようやくこれらの8割を理解できること，そして抽出はBNCの3,000～6,000語の頻度帯を中心に10,000語を越えるレベルにまで及ぶことから，「上級レベル」学習者に有効であると考えられる。

5．おわりに

本稿では，9種類の統計指標を利用して，100万語超からなる3種の話し言葉コーパスの特徴語を抽出し，検討した。9種の指標によって抽出された特徴語上位に現れた語の実例を参照し，出現度数による比較，単語の長さの比較，特徴語が属するBNC頻度帯の分布，特徴語の学年分布，学校英語教科書語彙の既習語の割合を調査した結果，各統計指標は100万語超という規模の話し言葉のコーパスにおいても，初級，中級，上級とそれぞれ特定の語彙レベル別に特徴のある語を抽出していることが検証された。従って，語彙の選定者が，学習対象者の習熟度を考慮し，これらの統計指標を上手に使い分ければ，効率的にESP分野の特徴語リストの作成が可能であることが確認できた。本研究の成果は，ESP分野の専門的知識を有しない英語教師にとっても，ESPの特徴語を選定する際の一助になると期待できる。

今後の課題は，実際にこれらの統計指標を適用して初級・中級・上級用のレベル別ESP学習語彙を選定し，語彙指導用Web教材を作成することである。

謝辞： 本研究は平成 17～18 年度科学研究費補助金（課題番号 17520401）を受けています。

参考文献

1）Hutchinson, T. and Waters, A. (1987) *English for Specific Purposes : A Learning-Centered Approach.* Cambridge : Cambridge University Press.

2）深山晶子，野口ジュディー，寺内一，笹島茂，神前陽子（2000）『ESP の理論と実践』東京：三修社.

3）Orr, T. (2002) "Twelve ESP Program Models for Study and Reflection." *JACET Summer Seminar Proceedings* (*New Perspectives in ESP*) 2 : 25-30.

4）Noguchi, J. (2002) "ESP : Where are we and where do we want to go?" *JACET Summer Seminar Proceedings*(*New Perspectives in ESP*) 2 : 18-24.

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

6）竹蓋幸生（1981）『コンピューターの見た現代英語』東京：エデュカ出版.

7）Bramki, D. and Williams, R. (1984) "Lexical Familiarization in Economics Text, and its Pedagogic Implications in Reading Comprehension." *Reading in a Foreign Language*, 2(1), 169-181.

8）Baker, M. (1988) "Sub-Technical Vocabulary and the ESP Teacher : An Analysis of Some Rhetorical Items in Medical Journal Articles." *Reading in a Foreign Language*, 4(2), 91-105.

9）Chujo, K. and Genung, M. (2004) "Comparing the Three Specialized Vocabularies Used in 'Business English,' TOEIC, and British National Corpus Spoken Business Communications." *Practical English Studies*, 11, 1-15.

10) Chujo, K. and Utiyama, M. (2006) "Selecting Level-Specific Specialized Vocabulary Using Statistical Measures." *System*, 34(2).

11) Thorndike, E.L. (1921) *The Teacher's Word Book*. New York : Teachers College, Columbia University.

12) Faucett, L. and Maki, I. (1932) *A Study of English Word-Values Statistically Determined from the Latest Extensive Word-Counts*. Tokyo : Shinozaki Shorin.

13) Thorndike, E.L. and Lorge, I. (1944) *The Teacher's Word Book of 30000 Words*. New York : Bureau of Publications, Teachers College, Columbia University.

14) 竹蓋幸生（1988）「キーワード 5000：SYSTEM について」『言語行動の研究』１，千葉大学英語学・言語行動研究会，88-93.

15) Dale, E. (1931) "Comparison of Two Word Lists." *Educational Research Bulletin*, 10, 484-489.

16) Dolch, E.W. (1936) "A Basic Sight Vocabulary." *Elementary School Journal*, 36, 456-460.

17) Johnson, D.D. (1971) "A Basic Sight Vocabulary for Beginning Reading." *Elementary School Journal*, 72, 29-34.

18) 清川英男（1976）「Spoken Word List に関する考察」『英語教育』25 (2)，42-46.

19) 竹蓋幸生（1981）前掲書.

20) 大学英語教育学会 (JACET) 教材研究委員会（1983）『「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』東京：大学英語教育学会.

21) 東京都中学校英語教育研究会研究部（1986）「英語基本語彙 1000 語，補足 460 語，外来語（英語）400 語」『語彙と英語教育(9)』.

22) 内山将夫，中條清美，山本英子，井佐原均（2004）「英語教育のための分野特徴単語の選定尺度の比較」『自然言語処理』11 (3)：165-197.

23) 竹蓋幸生（1981）前掲書.

24) 石川由紀，田中貴美枝，高橋秀夫，竹蓋幸生（1987）「ビジネス英語の語彙」『語学教育研究所紀要』１：53-66.

25) 竹蓋幸生，高橋秀夫，星野昭彦（1987）「計算機科学の語彙－コンピュータを英語で学ぶために」『千葉大学教育工学研究』８：27-40.

26) 中條清美，竹蓋幸生（1989）「女性向け英語雑誌の語彙」『時事英語学研究』28：73-84.

27) Scott, M. (1996/1999/2004) WordSmith Tools [Computer software]. Oxford : Oxford University Press.

28) Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge : Cambridge University Press.

29) Nelson, M. (2000) "A Corpus-Based Study of Business English and Business English Teaching Materials." Unpublished Ph. D. Thesis, Manchester : University of Manchester.

30) Flowerdew, L. (2003) "A Combined Corpus and Systemic-Functional Analysis of the Problem-Solution Pattern in a Student and Professional Corpus of Technical Writing." *TESOL Quarterly*, 37(3), 467-487.

31) Tribble, C. (2000) “Genres, Keywords, Teaching : towards a Pedagogic Account of the Language of Project Proposals.” In : Burnard, L. and T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main : Peter Lang Pub., 76–90.
32) Kennedy, G. (2003) “Amplifier Collocations in the British National Corpus : Implications for English Language Teaching.” *TESOL Quarterly*, 37(3), 467–487.
33) Oakes, M. (1998) *Statistics for Corpus Linguistics*. Edinburgh : Edinburgh University Press.
34) 小学館コーパスネットワーク BNC http://www.corpora.jp/
35) 中條清美，内山将夫(2004)「統計的指標を利用した特徴語抽出に関する研究」『関東甲信越英語教育学会研究紀要』18：99–108.
36) 内山他（2004），前掲論文.
37) 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.
38) Chujo, K. and Utiyama, M. (2005) “Selecting Level-Specific BNC Applied Science Vocabulary Using Statistical Measures.” *Selected Papers from the Fourteenth International Symposium on English Teaching*, English Teachers’ Association/ROC Taipei, 195–202.
39) Chujo, K. and Utiyama, M. (2006) 前掲論文.
40) 中條清美，内山将夫(2005)「語彙分析入門：lemma リストの作成」第 26 回英語コーパス学会ワークショップ，昭和女子大学，10/22/2005.
41) CLAWS7 (1996) http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html.
42) Chujo, K. (2004) “Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List.” In : J. Nakamura, N. Inoue, and T. Tabata (Eds.), *English Corpora under Japanese Eyes*, Amsterdam : Rodopi, 231–249.
43) Dale, E. and O’Rourke, J. (1981) *The Living Word Vocabulary*. Chicago : World Book–Childcraft International, Inc.
44) Harris, A.J. and Jacobson, M.D. (1972) *Basic Elementary Reading Vocabularies*. New York : The Macmilan Company.
45) 笠島準一他（2002）『New Horizon English Course 1，2，3』東京：東京書籍.
46) 市川泰男，安吉逸季，Hestand, J.R.，塩川春彦，小林千春，萩野敏(2003)『Unicorn English Course Ⅰ，Ⅱ，Reading』東京：文英堂.
47) 出版労連（1987）『教科書レポート』No. 31．出版労連.
48) 出版労連（2002）『教科書レポート』No. 46．出版労連.
49) Manning, C.D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge : The MIT Press.
50) Manning, C.D. and Schütze, H. (1999), 前掲論文.
51) Wakaki, M. and Hagita, N. (1996) “Recognition of Degraded Machine-Printed Characters Using a Complementary Similarity Measure and Error-Correction Learning.” *IEICE Trans. Inf. & Syst*. E79–D, 5.
52) Dunning, T.E. (1993) “Accurate Methods for the Statistics of Surprise and Coincidence.” *Computational Linguistics*, 19(1), 61–74.
53) Hisamitsu, T. and Niwa, Y. (2001) “Topic-Word Selection Based on Combinatorial Probability.” *NLPRS–2001*, 289–296.
54) Hisamitsu, T. and Niwa, Y. (2001), 前掲論文.
55) Manning, C.D. and Schütze, H. (1999), 前掲論文.
56) Rayner, J.C.W. and Best, D.J. (2001) *A Contingency Table Approach to Nonparametric Testing*. New York : Chapman & Hall/CRC.
57) 内山他（2004），前掲論文.
58) 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.
59) Schmitt, N. (2000) *Vocabulary in Language Teaching*. Cambridge : Cambridge University Press.
60) Handford, M. (2005) ‘A Copus-Based Interpretation of Spoken Business English’ Paper presented at Teaching and Learning of English : Towards an Asian Perspective (TLEiA) Conference 11/14–16/2005, Penang, Malaysia.
61) Íñigo-Mora, I. (2004) “On the Use of the Personal Pronoun We.” *Journal of Language and Politics* 3(1), 27–52. http://www.benjamins.com/jbp/series/JLP/3–1/art/0003a.pdf
62) McCarthy, M. (2002) “What is an Advanced Level Vocabulary?” in Tan, M. (ed) *Corpus Studies in Language Education*. Bangkok : IELE Press, pp. 15–29.
63) McCarthy, M. (2004) “Using Corpora to Understand Vocabulary, Collocation in Vocabulary Teaching and Learning.” Lecture presented at

The 32nd JACET Summer Seminar, Kusatsu Seminar House, 8/18-21/2004.

Appendix　使用した統計指標の定義式

	ESP リスト	BNC Spoken リスト
単語 α	a	b
単語 α 以外	c	d

a＝ESP リストに単語 α が出現した回数
b＝BNC Spoken リストに単語 α が出現した回数
c＝ESP リストの延べ語数－a
d＝BNC Spoken リストの延べ語数－b
n＝a＋b＋c＋d

$$LL_0 = a\log(an/((a+b)(a+c))) + b\log(bn/((a+b)(b+d))) + c\log(cn/((c+d)(a+c))) + d\log(dn/((c+d)(b+d)))$$

$$Chi2_0 = (n(ad-bc)^2)/((a+b)(c+d)(a+c)(b+d))$$

$$Yates_0 = n(|ad-bc|-n/2)^2/((a+b)(c+d)(a+c)(b+d))$$

Correction of the above three measures:

$$LL = sign(ad-bc)\times LL_0$$

$$Chi2 = sign(ad-bc)\times Chi2_0$$

$$Yates = sign(ad-bc)\times Yates_0$$

$$sign(z) = \begin{cases} +1 & if\ z>0 \\ -1 & otherwise \end{cases}$$

$$Dice = 2a/(2a+b+c)$$

$$Cosine = a/\sqrt{(a+b)(a+c)}$$

$$CSM = (ad-bc)/\sqrt{(a+c)(b+d)}$$

$$MI = \log(an/((a+b)(a+c)))$$

$$McNemar = \frac{(b-c)^2}{b+c}$$

$$Freq = a$$

注

注1）これらのESPリストからはBNC全体の頻度100以上(BNC HFWL)にない語を除外した．本研究の最終的な目標は，ESP語彙学習に最適な学習用語彙選定であるので，それらの低頻度語は効率的な語彙学習の観点からは必要性は低いと考えられた．

注2）各分野のテキストはmonologue（40%）とdialogue（60%）からなる．詳細は次の資料を参照されたい．Burnard, L. (2000) British National Corpus User Reference Guide, http://www.natcorp.ox.ac.uk/World/HTML/thebib.html

注3）BNCのspoken componentはEducational, Business, Public/Institutional, Leisure, Spoken Demographicの5分野からなる．①BNCのBusiness分野特徴語の抽出には，Business分野を除外した4分野の話し言葉コーパス5,838語（Educational＋Public/Institutional＋Leisure＋Spoken Demographic）を基準として使用した．②BNCのPublic/Institutional分野特徴語の抽出には，Public/Institutionalを除外した4分野の話し言葉コーパス5,610語（Educational＋Business＋Leisure＋Spoken Demographic）を基準として使用した．③BNCのLeisure分野特徴語の抽出には，Leisureを除外した4分野の話し言葉コーパス5,780語（Educational＋Business＋Public/Institutional＋Spoken Demographic）を基準として使用した．

注4）例えば，PMI（自己相互情報量）の場合，Business分野特徴語で155位まで，Public/Institutionalで390位まで，Leisureで252位までの指標値が同一値であった．

注5）機能語，内容語の定義はNation（2001：430-431）を参照した．

注6）The CANBEC corpus (Cambridge and Nottingham Business English Corpus): 1 million word database of transcribed business English http://www.nottingham.ac.uk/english/research/cral/projects.html

注7）Handford（2005）はpersonal inclusive *we*, personal exclusive *we*, corporate inclusive *we*, corporate exclusive *we*に注目した．さらに詳しく*we*の用法を知るにはÍñigo-Mora（2004）が参考になる．Two of the main uses of the personal pronoun “we” are the exclusive “we” and the inclusive “we”. Whereas the first one excludes the hearer (so “we”＝“I”＋my group), the second includes it (so “we”＝“I”＋“you”). Exclusive “we” represents a way of distancing, both from the hearer and from what the speaker is saying, and it is normally associated with power. (Íñigo-Mora, 2004: 34) なお，Quirk et al.（1985：350-351）では8種類の“we”の用法を区別している．

注8）ここでの語彙の比較は語の形式によるもので，意味別の比較は行なわれていない．実際には，抽出された特徴語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，意味も考慮すれば既習語の割合はより低くなる．なお，

用いた教科書は上級レベルの高等学校教科書である.

（H 18. 1 .10 受理）