JACET8000と大規模英語コーパス*

上　村　俊　彦

JACET8000 and Large English Corpora*

Toshihiko UEMURA

1．はじめに

大学英語教育学会の基礎語彙改訂委員会による語彙表（通称 JACET8000）は，BNC コーパスと同委員会による日本人学習者のためのコーパスをもとに日本の英語学習者向けの基本語彙リストとして確定された。基礎語彙表としての有効性については，すでに Mochizuki (2003), Murata et al. (2003), Uemura (2005), 上村 (2005) など，さまざまな検証がおこなわれてきたが，海外の大規模なコーパス，特にアメリカ英語コーパスや「話し言葉」コーパスによる検証が十分になされていなかった。本稿では，海外の大規模英語コーパス[1]と清水氏[2]によって独自に構築された大規模なアメリカ英語話し言葉コーパスを用いてJACET8000の有効性の検証，今後の拡張の可能性について検討する。

2．使用コーパスデータ

本稿で使用するコーパスデータは，American National Corpus 1st Release (ANC), British National Corpus (BNC), CNN Debate and Discussion 放送予稿スクリプト (CNN), International Corpus of English Great Britain (ICE-GB), Michigan Corpus of Academic Spoken English (MICASE) と Santa Barbara Corpus of Spoken American English Part-1 (SBC) である。[3] ちなみに，これらの大規模コーパスは，イギリス英語コーパス (BNC, ICE-GB) とアメリカ英語コーパス (ANC, CNN, MICASE, SBC) に分類できる。

3．6つのコーパスデータ

コーパスデータの語彙頻度を調べるために，清水氏開発の Perl スクリプト v8an を用いた。同スクリプトを使うと，インプットされたコーパスの語彙は，JACET8000 の語彙 (L1 to L8)，固有名詞・縮約形・非語の数 (others)，左記以外の語形 (over L9) に分類されて，個々の語形とその出現頻度が出力される。表1は，index[4]ベースで見た各コーパスの語数を示す。

全6コーパスの総 index 数は142,425，総 token 数は42,460,030であった。5コーパス全体における，L1 to L8, over L9, others, それぞれの index 合計の比率をグラフ化したものがグラフ1である。このグラフからも明らかなように，JACET8000 と一致する index の数は全 index の6%に留まる。

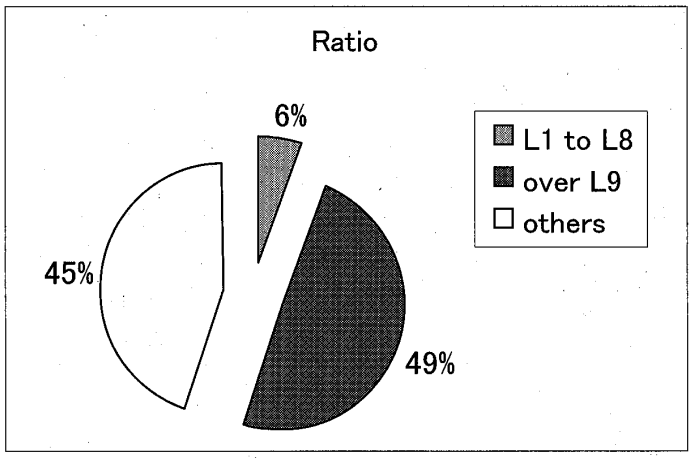
各コーパスによって出力される語形は，もととなった英文テキストデータのジャンルやトピッ

＊　本論文は，科学研究費補助金（課題番号16320076「大規模コーパスを用いた日本人英語学習者用の語彙リスト構築と教材分析システムの開発」平成16年～18年度）の助成による研究の成果の一部である。

表１　index 内訳

	L1 to L8	over L9	others
ANC	7321	8951	6002
BNC	8122	36328	24146
CNN	8071	38384	38770
ICE-GB	7219	6940	4046
MICASE	7383	15241	6224
SBC	4423	2509	1805

グラフ１　語形比率



クによってその出現頻度に大きなばらつきが生じる。コーパスの規模が大きくなるにつれて，JACET8000 以外の語形（over L9）とともに，固有名詞，数詞，縮約形（others）がコーパス全体に占める比率が大きくなる。頻度順語彙リストから，数詞，固有名詞，縮約形，非単語（index ベースで64,065語，token ベースで2,315,236語）を除外すると頻度順語彙リストの index 数は79,271となった。表２は，各 index の出現頻度による度数分布を示す。出現頻度100以下の index が総 index 数に占める割合は90.54％であった。出現頻度100回を超える index は全体の約10％で7497語であった。

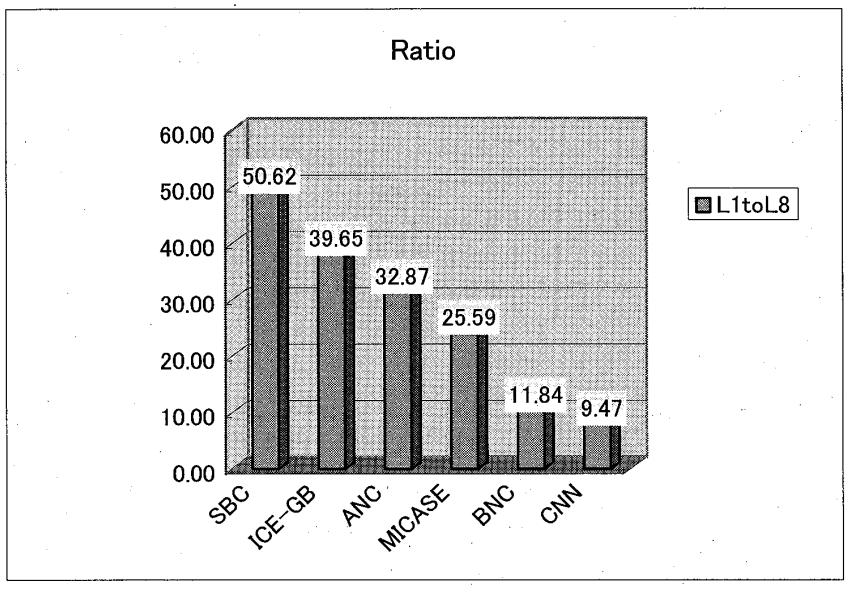
表２　出現頻度順 index 分布

出現頻度	index 数
１～10	58,610
10～100	13,164
100～1,000	5,249
1,000～10,000	1,857
10,000～100,000	386
100,000～1,878,034	5

すでに見たように，表１は各コーパスの規模を index ベースで示している。グラフ２は，JACET8000 の語形と一致する index 数が各コーパスの総 index 数に占める割合を示す。６つの

コーパスの中で SBC は最も小規模なコーパスであるが，JACET8000 と一致する index 数の全 index に占める比率はこの SBC が最も高い。表 1 とグラフ 2 からも明らかに，JACET8000 に一致する index 数の総 index 数に対する割合とコーパスの規模とは反比例の関係にある。

グラフ 2　各コーパスデータに占める JACET8000 語形（総 index 数）の割合

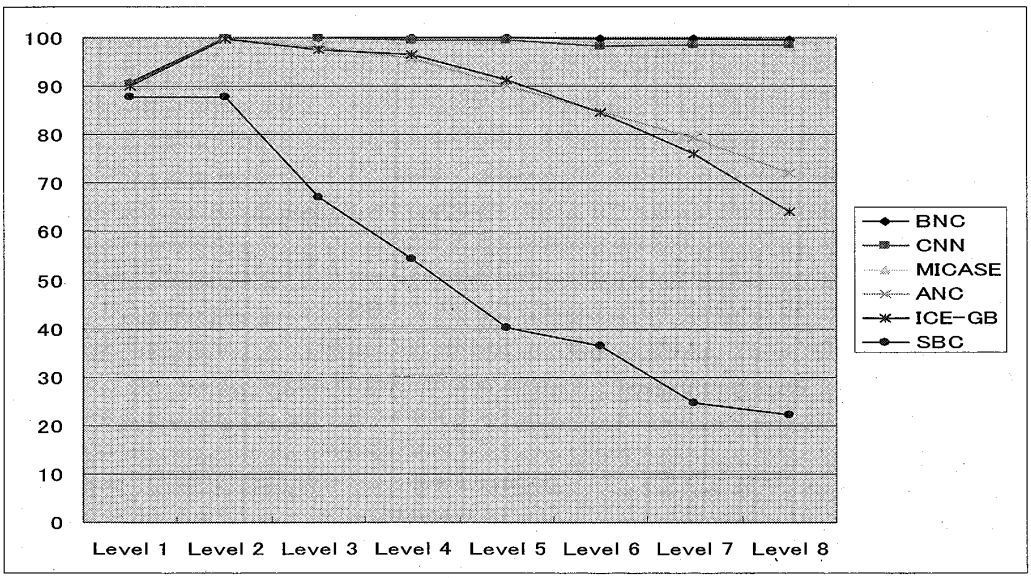


4．JACET8000 と 6 つのコーパス

JACET8000 は頻度順にレベル 1 からレベル 8 までの 8 レベルによって構成されている。英文テキストの語彙レベルを判定する清水 v8an は，レベルごとの一致した語形の出現頻度を出力する。ただし，レベル 1 は本来の1000語に Plus250 を加えた合計1250語となっている。

v8an を使って 6 つのコーパスの語彙について，レベルごとのカバー率をグラフ化したものがグラフ 3 である。

グラフ 3　JACET8000 レベルごとの出現率



BNC と CNN については，レベル1を除くとほぼ100%のカバー率となった。残り4つのコーパスについて，カバー率を見ると，SBC のみがほか3つ（MICASE, ANC, ICE-GB）と大きく異なった。

表3　Pearson 相関

	BNC	CNN	MICASE	ANC	ICE-GB	SBC
BNC		0.984				
CNN	0.984					
MICASE				0.998	0.986	0.781
ANC			0.998		0.989	0.797
ICE-GB			0.986	0.989		0.766
SBC			0.781	0.797	0.766	

コーパス相互の相関をみた表3からも明らかなように第1群の BNC と CNN，また第2群の ANC，ICE-GB と MICASE の相互間には有意基準1%で強い相関が認められた。また，第3群の SBC と第2群 ANC, ICE-GB, MICASE の各コーパス間には，5%基準で有意な相関関係が認められた。

SBC は，International Corpus of English（以下，ICE）のアメリカ英語版として構築された英語コーパスの一部である。コーパスの元となったのは，さまざまな地域，職業，社会，あるいは人種的な背景をもったアメリカの人々がおこなった，うわさ話，議論，セールス・トーク，職場・市議会における会話，講義，政治演説，説教などを収録した話し言葉の音声データである。今回使用したものはこの音声データを電算処理可能な文字データ化したもので，6つのコーパスデータの中で最も小規模なものである。相関係数の数値から判断すると，SBC は同じく ICE コーパスであるイギリス英語コーパス ICE-GB よりも，アメリカ英語コーパスである ANC やアカデミックな話し言葉からなる MICASE との相関性が若干高い結果となった。

SBC を除いた5つのコーパスでは，レベル2のカバー率がレベル1のカバー率を大きく上回る結果となった。この現象は，v8an のレベル1に本来の出現頻度に係わらず日本における英語の導入教育で基本語と考えられている語彙250語（Plus250）を加えたために生じたものと考えられる。ちなみに，レベル2の1000語に限定すると，5つのコーパスではほぼ100%の一致となる。

5．JACET8000の拡張

すでに表2で見たように，6つのコーパスデータ全体で出現頻度が100回以下の index 数が全 index 数に占める割合は90.54%であったが，全語形に占める比率は低い。index 数は少ないが全語形に占める割合が高い頻度100回以上の index のうち，v8an によって over L9（数詞，固有名詞，縮約形，非単語を除外した JACET8000 に含まれない語形）として出力されたものは996語であった。この中で，出現頻度1000回以上の語形は以下の通りであった。

Er (86407)	erm (46664)	gonna (17650)	uh-huh (16092)	um-hum (15116)
Cos (11668)	ph (10151)	uhm (5393)	xx (5321)	wanna (4940)
Mhm (4750)	gotta (3662)	th (3628)	al (3289)	jack (3066)

forth (2700)	re (2468)	aha (2425)	cuz (2194)	ve (2178)
Don (1921)	innit (1904)	email (1877)	courtroom (1778)	Graham (1708)
hm (1486)	DNA (1480)	s (1417)	juror (1411)	Dunno (1405)
anthrax (1312)	cooperate (1258)	rick (1193)	cooperation (1193)	so-called (1190)
th- (1073)	de (1037)	capitol (1027)	scary (1021)	insurgent (1019)
ll (1018)	miller (1007)	da (1005)		

この中には，v8anで排除できなかった固有名詞（Jack, Don, Graham, Miller, Rick, Capitol），縮約形の一部（don, ve, s, ll）非語（xx, th, hm）接辞（re），頭字語（pH, DNA），分かち書きされた外来語の一部（al, de, da）が含まれている。

6つのコーパスは話し言葉のデータも多く含まれているためか，間投詞（er, uh-huh, uhm, um-hum, aha, hm），話し言葉で特徴的な異綴り（Cos, cuz→(be)cause; gonna→going; wanna→want; gotta→(have) got to; Dunno→don't know）が多用されている。

以上のものを除くと，出現頻度1000回以上で，JACET8000のリストに漏れた語形の中で追加語として検討すべきものは，forth, email, courtroom, juror, anthrax, cooperate, cooperation, so-called, scary, insurgentとなる。

上記の追加候補語は，従来のJACET8000に不足していたアメリカ英語コーパス，話し言葉コーパスの充実を反映した大容量のコーパスデータに基づくものであるが，依然としてコーパスのベースとなった英文テキストのジャンルの傾向が反映している。たとえば，courtroom, jurorは法律関係の英文テキスト，anthraxは時事的な英文テキストの偏重が推測される。各コーパスは元となったジャンルによる偏りが顕著化される傾向があるため，コーパス・レンジ(出現したコーパス数）によりコーパス間のバランスを取ることが必要となる。以下，4つのアメリカ英語コーパスについて，JACET8000の追加候補語の抽出をおこなう。

5．1　MICASE

表4はv8anでover L9として出力されたMICSEの出現頻度100回以上の語形について，その出現頻度，ほかの5つのコーパスにおける総出現回数，log likelihood（LL）値[5]，コーパス・レンジ値を調べたものである。

表4　MICASEによる追加候補語

LV	WORD	TGT	BASE	LL	RANGE
L09->	gonna	4192	13458	8337.57860	6
L09->	cuz	2192	2	13838.89518	2
L09->	wanna	1983	2957	6148.55672	5
L09->	gotta	318	3344	139.28836	5
L09->	email	267	1610	293.10177	6
L09->	semester	192	128	795.2797774	5
L09->	forth	164	2536	20.16575	6
L09->	photon	156	14	891.63611	3
L09->	serotonin	153	19	850.271608	2
L09->	lotta	145	9	849.74182	3
L09->	precipitate	137	47	661.8711183	4

L09->	hafta	126	0		1
L09->	sorta	125	4	755.66813	3
L09->	handout	105	156	326.10637	5

頻度100回以上の語形のうち，テキスト間のばらつきを最小とするためにレンジ（4コーパス以上），LL値（0～1000）を追加候補語抽出の基準とすると，MICSE では email, semester, forth, precipitate, handout が候補となる。なお，JACET8000 の現在の「語の定義」を踏襲すると，話し言葉の異綴り（gonna, cuz, wanna, gotta, lotta, hafta, sorta）は追加語とならない。

5．2　ANC

ANC について，MICASE と同じ条件で追加語を抽出したものが表5である。

表5　ANC による追加候補語

LV	WORD	TGT	BASE	LL	RANGE
L09->	forth	369	2331	107.9335	6
L09->	scary	317	704	467.5243	6
L09->	trash	232	363	447.5069	5
L09->	aerobics	136	69	443.6126	5
L09->	mow	129	45	466.9496	5
L09->	awhile	127	300	177.2085	4
L09->	anytime	116	356	123.5279	5

表5の語形については，すべてを ANC による追加候補語とする。

5．3　SBC

追加語候補の基準に該当する index は存在しなかった。

5．4　CNN

CNN について，MICASE と同じ条件で追加語の抽出をおこなうと，該当の index 数は280語となった。以下の表6は，紙面の都合でレンジ5までの追加候補の語形である。

表6　CNN による追加候補語

LV	WORD	TGT	BASE	LL	RANGE
L09->	forth	1868	832	114.5317	6
L09->	email	1565	312	510.3748	6
L09->	cooperate	1178	80	781.7707	6
L09->	so-called	994	196	328.0656	6
L09->	cooperation	911	282	155.4229	6
L09->	scary	617	404	0.609486	6
L09->	makeup	507	104	160.1652	6
L09->	pacific	483	92	165.3773	6
L09->	heck	461	249	9.708347	6

L09->	journalistic	458	12	389.5291	6
L09->	skeptical	378	113	68.96697	6
L09->	well-known	310	64	97.21689	6
L09->	zoo	297	58	99.04936	6
L09->	tremendously	296	98	43.88147	6
L09->	lynch	290	21	187.6839	6
L09->	fingerprint	290	39	134.219	6
L09->	perpetrator	276	14	202.0326	6
L09->	darn	244	117	10.72779	6
L09->	closet	229	91	21.05695	6
L09->	someday	228	69	40.64504	6
L09->	coma	206	35	78.86293	6
L09->	stabilize	192	51	42.92146	6
L09->	brad	191	33	71.99598	6
L09->	abide	185	57	31.854	6
L09->	paralyze	181	75	14.52104	6
L09->	mobilize	155	55	19.53226	6
L09->	casino	152	28	53.8524	6
L09->	truthful	142	38	31.37217	6
L09->	rainy	134	70	3.589189	6
L09->	adorable	130	16	63.94472	6
L09->	relive	129	12	74.56949	6
L09->	alluded	128	21	50.56771	6
L09->	collateral	114	9	71.16156	6
L09->	graduation	114	32	23.27763	6
L09->	harden	114	64	1.731566	6
L09->	spike	114	74	0.154497	6
L09->	paranoid	110	59	2.443144	6
L09->	old-fashioned	106	50	5.049481	6
L09->	autograph	103	13	49.82228	6
L09->	juror	1345	66	994.1054	5
L09->	columnist	908	13	837.7876	5
L09->	recount	870	25	727.9171	5
L09->	videotape	718	48	479.3328	5
L09->	carol	675	218	105.5847	5
L09->	ranch	633	69	335.714	5
L09->	long-term	612	222	72.91753	5
L09->	rabbi	516	47	301.4303	5
L09->	pastor	480	16	390.1033	5
L09->	max	456	142	76.87926	5
L09->	audition	424	38	249.6631	5
L09->	syndicate	401	50	195.5546	5

L09->	excerpt	397	17	304.4864	5
L09->	bestseller	381	9	329.5825	5
L09->	gag	361	20	257.1918	5
L09->	wrestling	342	73	103.1629	5
L09->	believer	337	58	127.4886	5
L09->	hijack	335	27	207.2272	5
L09->	eyewitness	330	12	263.208	5
L09->	abduct	310	15	230.0451	5
L09->	spokesperson	305	34	159.6293	5
L09->	anytime	302	170	4.473314	5
L09->	phenomenal	291	86	54.23685	5
L09->	militarily	287	13	216.8976	5
L09->	cookie	265	72	57.13225	5
L09->	hype	261	37	116.1922	5
L09->	nightly	258	17	173.1712	5
L09->	devastation	252	14	179.3767	5
L09->	cooperative	251	135	5.459875	5
L09->	follow-up	248	34	113.1899	5
L09->	inaugural	243	14	170.9405	5
L09->	resonate	234	12	170.7333	5
L09->	rehearse	230	66	45.20288	5
L09->	wade	225	34	95.28268	5
L09->	tornado	225	106	10.78189	5
L09->	fundraising	217	16	139.415	5
L09->	overly	208	37	76.26688	5
L09->	insanity	206	22	110.5557	5
L09->	accomplishment	206	25	102.2619	5
L09->	fireman	201	107	4.719209	5
L09->	clone	199	52	45.6386	5
L09->	escalate	198	70	25.17997	5
L09->	dismantle	195	39	63.36979	5
L09->	lobbyist	180	22	88.94402	5
L09->	implant	173	34	57.30248	5
L09->	upbeat	170	14	104.152	5
L09->	backyard	170	106	0.641779	5
L09->	respectful	168	20	84.43483	5
L09->	steroid	168	29	63.37961	5
L09->	abusive	168	35	52.1388	5
L09->	northwest	167	38	46.54282	5
L09->	Valentine	166	49	31.00712	5
L09->	short-term	161	99	0.785413	5
L09->	uproar	160	21	75.34074	5

L09->	surrogate	159	23	69.69865	5
L09->	nuance	153	24	62.8265	5
L09->	demeanor	150	9	104.0914	5
L09->	envision	147	33	41.68168	5
L09->	ramifications	144	23	58.23071	5
L09->	moderation	143	20	64.31089	5
L09->	publicize	143	66	7.532472	5
L09->	sideline	138	23	53.75977	5
L09->	intentionally	136	36	30.56903	5
L09->	spirituality	135	22	53.64932	5
L09->	clout	135	37	28.6928	5
L09->	nurture	134	37	28.13144	5
L09->	unconstitutional	132	14	71.12194	5
L09->	blackout	131	24	46.66641	5
L09->	waive	131	31	34.66951	5
L09->	inspirational	128	7	91.56853	5
L09->	undercover	127	18	56.54706	5
L09->	southwest	127	54	9.304851	5
L09->	pajama	127	82	0.204322	5
L09->	breakup	126	8	85.72301	5
L09->	privy	126	13	68.95072	5
L09->	tyranny	126	18	55.76573	5
L09->	loophole	126	36	24.95274	5
L09->	commend	126	50	11.63632	5
L09->	heartbeat	125	18	54.98638	5
L09->	brother-in-law	125	67	2.790419	5
L09->	wrestle	122	26	36.86967	5
L09->	fetus	121	30	30.00781	5
L09->	embargo	121	40	17.9982	5
L09->	discredit	119	7	83.13464	5
L09->	obese	119	19	48.13628	5
L09->	leverage	118	31	26.83828	5
L09->	intrigue	118	62	3.038521	5
L09->	melanoma	117	21	42.52915	5
L09->	fisher	117	39	17.08598	5
L09->	repeal	115	38	17.1221	5
L09->	genocide	113	7	77.55347	5
L09->	stepmother	113	8	73.79305	5
L09->	inclusive	112	35	18.7502	5
L09->	safeguard	112	63	1.670487	5
L09->	unanswered	111	14	53.71705	5
L09->	vaccination	111	18	44.30189	5

L09->	lineup	110	23	33.99532	5
L09->	astound	110	29	24.8822	5
L09->	psych	110	73	0.05843	5
L09->	disconnect	109	18	42.81329	5
L09->	reorganize	109	33	19.41639	5
L09->	smuggle	108	24	31.01674	5
L09->	prewar	107	32	19.50727	5
L09->	dime	106	29	22.59554	5
L09->	rehabilitate	104	44	7.758172	5
L09->	lousy	103	57	1.779231	5
L09->	brutality	101	11	53.59271	5
L09->	repercussion	101	20	33.18175	5
L09->	nighttime	101	33	15.41144	5
L09->	adrenaline	101	38	10.90858	5
L09->	hesitant	100	26	23.11137	5

上記のうち，固有名詞か普通名詞か不明な carol（675），pastor（480），Valentine（166），fisher（117）については，本来の英文テキストに当たり追加候補語とするかどうかの再検討が必要である。JACET8000 では，「語の定義」により省略語を排除しているため，max（456）については採用しない。

CNN による追加候補語は多い。これは，CNN コーパスの規模が他3つよりも大幅に大きいこと，CNN コーパスのソースとなった TV 討論番組のトピックの多様性が起因している。ちなみに，総出現頻度1751回の courtroom は，LL 値を除く他のすべての要件を満たしているが LL 値（1602.758）のために表6に記載されていない。

なお，表6は元のリストを，レンジ，出現頻度，LL 値の順にソートして配列した非レマリストであり，JACET8000 にレマ化された語形が存在するものも含まれている。また，JACET8000 では排除された動詞原形以外の分詞形，名詞や形容詞における接辞，複合語などの語形も含まれている。これらに該当する個々の語形については，今後，出現頻度，LL 値，レンジを再度検討するとともに今回と同規模のコーパスによる検証をおこない追加語とするかどうか決定すべきである。

6．終わりに

JACET8000 の語形について，5つの海外大規模英語コーパスと大規模 CNN 話し言葉コーパスを用いたテキストカバー率の検証をおこなった。index ベースでは，全語彙データの6％程度にすぎない JACET8000 であるが，各コーパスにおける出現頻度100回以上の語形については高い一致率となった。

JACET8000 の拡張のために，アメリカ英語コーパスや話し言葉コーパスから追加語候補を抽出する作業をおこなった。アメリカ英語の話し言葉コーパス4種類について，追加候補語の抽出をおこなったところ，SBC（0），MICASE（5），ANC（7），CNN（271）となった。CNN が大きくほかのコーパスを引き離した数値となったのは，index ベースで CNN は次に規模が大きい MICASE の2倍規模のコーパスであること，コーパスの元となった CNN 討論番組のジャンルが多岐にわたることに起因していることが予想される。

注

1）海外の大規模コーパスの購入に際しては，2004年度から3年間に渡る課題研究に対して助成された科学研究費補助金により購入した。

2）JACET8000メンバーの1人，清水伸一氏（安城学園高校）が収集したCNN番組Interview and Debateカテゴリの Larry King Live, On the Story, Late Edition, Reliable Sourcesの番組スクリプト（2000年1月－2006年6月放送分）。

3）以下は各コーパスの概要（アルファベット順）

（ANC）BNCのアメリカ版。コーパス全体が11,508,216語で構成。「話し言葉」データ3,224,388語，「書き言葉」データ 8,283,828語。
http://americannationalcorpus.org/FirstRelease/contents.html

（BNC）20世紀イギリス英語の1億語コーパス。全体は「書き言葉」コーパス90%，「話し言葉」コーパス10%の構成。データ収集時期　1991年～1994年。本研究では，「話し言葉」コーパスのみ使用。
http://www.natcorp.ox.ac.uk/corpus/index.xml

（CNN）清水氏によるCNN番組 Interview and Debate（2000年1月～2006年6月放送の Larry King Live, CNN Late Edition と Reliable Sources）のスクリプト。index 85,225，token 25,150,021。

（ICE-GB）世界のさまざまな英語の比較検討のために開始された英語コーパス構築プロジェクトの最初の成果。現代イギリス英語の話し言葉と書き言葉100万語からなるコーパス。
http://www.ucl.ac.uk/english-usage/ice/

（MICASE）現代の学術的な話し言葉（contemporary academic speech）を1997年から2001年にかけて収集。音声データから書き取った152のトランスクリプト，総語数 1,848,364語によって構成されている。
http://micase.umdl.umich.edu/m/micase/

（SBC）249,000語からなる現代アメリカ英語の話し言葉コーパス。ICE プロジェクトのアメリカ英語の一部。
http://projects.ldc.upenn.edu/SBCSAE/

4）Perl言語（Active Perl）上で清水氏のスクリプト v8an.pl を稼働させ，6つの大規模英語コーパスのテキストデータ処理をおこなった。index は統語論・語源学上関係のある複数の語形を1つの代表的な語形にまとめたもので，清水氏によると WordSmith における type よりも lemma に近い。

5）LLについては以下の記述を参照。

"The word frequency list is then sorted by the resulting LL values. This gives the effect of placing the largest LL value at the top of the list representing the word which has the most significant relative frequency difference between the two corpora.... the words most indicative (or characteristic) of one corpus, as compared to the other corpus, at the top of the list. The words which appear with roughly similar relative frequencies in the two corpora appear lower down the list." (Rayson & Garside (2000))

参照文献

（和　文）

上村俊彦(2005)「JACET8000 と電子版ニューヨーク・タイムズ紙掲載の新刊書第１章の語彙」『県立長崎シーボルト大学国際情報学部紀要』pp.263-272.

（英　文）

Kilgarriff, A. (2005) "Language is never ever ever random" in *Corpus Linguistics and Linguistic Theory* 1 (2): 263-276.

---------- (2001) "Comparing Corpora." in *International Journal of Corpus Linguistics* 6 (1): 1-37.

Leech, G.; Rayson, P.; & Wilson, A. (2001) *Word Frequencies in Written and Spoken English based on the British National Corpus.* Harlow: Pearson Education Limited.

Mochizuki, M. (2003) "JACET8000 Compared with Other Vocabulary Lists." *ASIALEX'03 Tokyo Proceedings: Dictionaries and Language Learning: How can Dictionaries Help Human &* Machine Learning? pp. 378-383.

Murata, M.; Shimizu, S.; Sugimori, N.; Ishikawa, S. & Mochizuki, M. (2003) "JACET8000: a word list constructed using a scientific method and its applications to language teaching and learning" In Murata, M. et al. eds. *ASIALEX'03 Tokyo Proceedings.* pp. 356-378.

Rayson, P. and Garside, R. (2000). "Comparing corpora using frequency profiling." In Proceedings of the *Workshop on Comparing Corpora, Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000).* 1-8 October 2000, Hong Kong, pp. 1-6.

Uemura, T. (2005) "How Good Are Graded English Readers for ESL/EFL Students?" *Proceedings of the 4th Asialex Conference* pp.338-344.