English Corpus Studies, 第 12 号, pp.19-35, 2005, 中條・内山・長谷川

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

A Study on Extracting Domain-Specific Vocabulary from Current English Texts Using Statistical Measures

This study explores the effectiveness of using statistical measures to identify specialized vocabulary in English texts widely available on the Internet and which could be developed as authentic resources for pedagogical purposes. Chujo and Utiyama (2004) established these measures as effective for identifying vocabulary from a 100,000-word corpus. The focus of this study has been to determine the effectiveness for smaller (20,000-word) corpora. Eight statistical measures were applied to 18 twenty-thousand-word mini-corpora to identify its domain-specific words, and an examination of the resulting vocabulary lists showed 1) that each statistical measure extracted a different level of domain-specific words by its word length, vocabulary level, grade level, and high school textbook vocabulary coverage; 2) that specific measures produced level-specific words, i.e., frequency, *Dice coefficient* and *complimentary similarity measure* identified beginning-level words, *cosine / log-likelihood ratio / chi-square test / chi-square test with Yates correction* produced intermediate-level words, and *mutual information* created lists of advanced level words. We can conclude that the application of statistical measures can be an effective tool for identifying and selecting vocabulary for pedagogical purposes.

1. はじめに

インターネットの普及とともに，世界の最新情報を瞬時にダウンロードして時事英語の教材として利用することが容易になった。本物の教材(authentic material)は学習者の興味と関心も高く，英語学習のやる気を喚起することに役立つ(Rixson,1998)。しかし同時に，時事英語の問題として学習者の語彙不足が深刻な問題であるとの指摘がなされている(東,1998)。大学はもとより高等学校においても，英字新聞は NIE(Newspaper in Education)の一環として格好の学習素材であるが，時事用語の存在がその学習を難しくしている(谷口, 1998)。そこで，学習者と教材の語彙レベルを乖離させないように，[1] 学習者の習熟度に応じて，適切な語彙の補充が必要になる。[2]

教育用語彙選定の分野においては，従来，頻度や分布度に選定者の主観を加えた手法を用いて語彙選定を行なっていたが，近年，コーパス言語学や自然言語処理の発展とともに，統計的指標を取り入れる傾向がある(Oakes,1998)。例として，大学英語教育学会の JACET8000(2003)では，British National Corpus(以

下 BNC と表記)の頻度データに，対数尤度比という統計値を用いて単語の頻度順位を補正する手法が用いられた。対数尤度比などの統計的指標は英文資料に出現する単語どうしの共起頻度を利用して類似度を計算し，各単語に独自の値を与えるもので，値の順位付けを利用してコーパスの分析に有効な情報を得られることが知られている(齊藤他，1998：196)。例えば，英文検索プログラムとして英語教師に普及している WordSmith には，[3] 単語によるテキストの特徴度を示すために対数尤度比とイエーツの補正公式を利用した機能が付加され，小学館 BNC オンラインサービスでは，[4] 自己相互情報量と T-スコアを用いた共起語検索が利用できる。

しかし，種々の統計的指標が具体的にどのような単語に高い指標値を与えて「特徴度の高い単語」(以下，特徴語と表記)としているのか，どの指標とどの指標が似ているのか，どのように「良い」指標であるかということは，これまで明らかにされていなかった。

これに対し中條他(2004)，内山他(2004)では 9 種類の統計的指標(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 万語の言語資料 1 種類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討を行なった。その結果，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ことを確認し，これらの統計的指標の教育用語彙選定への利用可能性を示した。

本研究では，さらに実際の教育現場での実践的利用に向けて，比較的小規模な時事英語資料 18 種各 2 万語の特徴語抽出に統計的指標を適用し,特徴語を順位付けて抽出した実例を検討，比較することによって，語彙選定における統計的指標の実用的な有効性を考察することにした。このような実用的観点からの統計的指標の詳細な検討･比較の試みは，語彙選定のみならず，コーパス言語学や自然言語処理の分野においても有用であると考える。

2．研究の方法

2.1　言語資料

①　時事英語資料

言語資料には，**表 1** の各 2 万語の時事英語素材から作成した 18 種類の「時

事英語リスト」を用いた。**表 1** の異語数と延べ語数は，単語の屈折形とその頻度を基本形に集約し，固有名詞と数字等を除外した後の語数である。[5]

表 1　調査した時事英語資料

素 材 名	記 事 名	異語数	延べ語数
音声英語			
ABC News	Feared Dead, Found Alive 他 19 編	2,602	18,592
CBS News	Anti-Terror 'TIPS' Program On Hold 他 45 編	2,645	18,768
CNN News	Good Sign, Separated Baby Blinks 他 8 編	2,568	17,722
PBS News	Inspecting Iraq 他 9 編	2,163	19,239
BBC News	California Gets Landmark Green Law 他 13 編	2,935	18,278
VOA News	Explosion Rocks Tel Aviv Street 他 23 編	2,324	17,388
VOA Special English	Science Report: New Stem Cell Study 他 29 編	1,672	18,460
やさしいビジネス英語	Business Casual 等 5ヶ月分	2,402	19,973
文字英語			
Time	A Bad Menu for Peace 他 30 編	3,138	18,300
Newsweek	The Economic Blame Game 他 13 編	3,036	17,906
New York Times	U.S. Military Plane Crashes in Puerto Rico 他 18 編	2,583	18,052
USA TODAY	5 Shot Dead at Dallas Home 他 24 編	2,966	17,904
Chicago Tribune	Former Enron Exec Pleads Guilty 他 39 編	2,495	17,088
Japan Times	24-hour Party People 他 18 編	3,169	18,516
Daily Yomiuri	Stock Slump Deflates U.S. Bubble Economy 他 31 編	2,449	18,290
Asahi Weekly	Panel Calls for Cut in NTT Connection Fees 他 53 編	2,743	17,549
Student Times	The Segregation of Singaporean Education 他 19 編	2,900	18,519
News For You	Enron Under Investigation 等 10週分	2,118	17,580

② 基準言語資料

各時事英語リストに出現する語の生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，BNC で頻度 100 以上に該当する 13,994 語の「BNC リスト」(Chujo, 2004)を使用した。

③ 参考資料

抽出された特徴語の語彙レベルの調査には以下の 3 種の資料を使用した。

1) BNC リスト：頻度分布を調査するため，上記②の資料を用いた。

2) *The Living Word Vocabulary* ：学年分布を見るため，40,400 項目の語の意味に関して，米国人の 75%以上の子供が理解できる学年を調査した Dale & O'Rourke(1981)を用いた。

3) 中学校・高等学校教科書語彙リスト：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。これらの教科書は，一般的な傾向が得られるように，両者とも全国での採択率の高いものから選択した。

2.2 統計的指標

① 8 種の統計的指標

本研究では，英文の特徴を高く反映する特徴語を抽出できることが確認されている，頻度(竹蓋, 1981; Nation, 2001)，ダイス係数(Manning et al.,1999)，補完類似度(山本他, 2002), コサイン(Manning et al.,1999), 対数尤度比(Dunning,1993), イエーツの補正公式(Hisamitsu et al.,2001)，カイ二乗値(Hisamitsu et al.,2001), 自己相互情報量(Manning et al.,1999)の合計 8 種の統計的指標を用いた。単一の独立した指標である頻度に対し，対数尤度比，イエーツの補正公式，カイ二乗値，自己相互情報量は 2 つの確率変数間の比較にもとづく指標であり，ダイス係数,補完類似度,コサインは 2 つのベクトル間の比較にもとづく指標である。これらの指標は以下のパラメタ a,b,c,d によって計算される。

表 2　単語の出現状況を示すパラメタ

	時事英語リスト	BNCリスト
単語α	a	b
単語α以外	c	d

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

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の時事英語資料における出現状況が,一般分野の英文資料である BNC における出現状況よりも顕著であるならば，その単語は時事英語資料において特徴的な単語であろうという期待がある。[6] 各統計的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る。**表 2** の a,b,c,d を用いて，たとえば自己相互情報量は，

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

で求められる。その他の統計的指標も**表 2** のパラメタを用いて **Appendix 1** に示した定義式によって求められる。[7] 各統計的指標によって定義式における a,b,c,d の使い方が異なるため，同一の単語αであっても異なる指標値が与えられる。[8]

② 特徴語リストの作成

8 種類の各統計的指標を用いて，それぞれの時事英語資料における各単語の

出現状況を BNC リストでの出現状況と比較した指標値を求め,その指標値で降順にソートして特徴語リスト(8 指標×18 資料)を作成した。[9]

③　指標間の相関

各統計的指標がどの程度似た尺度であるかを調べるために，任意の 2 指標間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た。

2.3　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の比較には，まずマクロ的視点から，抽出された最上位 20 語についての出現度数の中央値，平均文字数の比較，規模を拡大した上位 300 語の累計的平均文字数の比較，学習語彙，BNC 頻度，学年分布から見た語彙レベルの比較を行ない，次にミクロ的視点から，一単位資料中で抽出された特徴語の具体例を比較・検討する方法を用いた。

3. 結果と考察

3.1　統計的指標によって抽出された特徴語の順位相関

各指標間の関係を調べるため，各時事英語資料について，任意の 2 指標間の特徴語リストの順位相関係数を計算し，それを 18 種について平均した結果を**表 3** に示した。[10]

表 3　指標間の順位相関

	対数尤度比	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	コサイン	ダイス係数	頻度
対数尤度比	-							
イエーツの補正公式	0.9	-						
カイ二乗値	0.9	0.9	-					
補完類似度	0.8	0.8	0.8	-				
自己相互情報量	0.8	0.7	0.8	0.7	-			
コサイン	0.7	0.7	0.7	0.6	0.6	-		
ダイス係数	0.5	0.5	0.5	0.7	0.4	0.7	-	
頻度	0.2	0.2	0.1	0.4	0.0	0.4	0.7	-

順位相関が0.9以上　順位相関が0.8以上0.9未満

表 3 より，対数尤度比，イエーツの補正公式，カイ二乗値は互いの順位相関が 0.9 以上であることから，良く似た順位付けで特徴語を抽出する指標であると言える。その他の指標はある程度独立した指標であるが,どちらかと言えば，補完類似度，自己相互情報量，コサインは前述の 3 指標に似ており，ダイス係数と頻度はあまり似ていないことがわかる。順位相関は各資料の特徴語全体の

順位についての関係を大局的に見たものであるが，実際に特徴語リストを利用する際にはリスト上位に順位付けられた指標値の高い単語に注目することが多い。次節では，上位に順位付けられた特徴語の例を見る。

3.2 上位 20 位に順位付けられた特徴語の例

18 種の時事英語資料すべてについて特徴語リストを作成したが，全リストへの言及は紙幅の関係で不可能であるため，本稿では，例として，中條他(2003)の時事英語資料の調査において中程度の語彙レベルと判定され，教材としてもよく使用される ABC News(音声言語)と *Japan Times*(文字言語)を検討する。[11] **表 4** と**表 5** に各指標より求められた ABC News，および，*Japan Times* の特徴語を指標値の高いものから降順に上位 20 語を示した。単語の右側の数値は ABC News，*Japan Times* の各 2 万語での出現度数，最下段は 20 語の出現度数の中央値と単語の平均文字数を示す。各特徴語の指標値の表示は省略した。カイ二乗値とイエーツの補正公式は上位 20 位に同じ語が現れたため，カイ二乗値のリストで代表した。[12] 今回調査した 2 万語の ABC News と *Japan Times* はそれぞれバラエティに富む内容の 20 編のニュースと 19 編の記事で構成されており，**表 4** と**表 5** の特徴語には多様な話題を想起させられる語が抽出されている。上位 20 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。

表 4 指標別 ABC News の特徴語上位 20 語

順位	頻度		補完類似度		ダイス係数		コサイン		対数尤度比		カイ二乗値 イエーツの補正公式		自己相互情報量	
1	the	1060	she	342	embryo	40	embryo	40	embryo	40	embryo	40	songwriter	5
2	be	800	say	242	one	69	teen	16	say	242	teen	16	smuggler	7
3	a	642	they	307	family	56	lure	13	she	342	lure	13	embryo	40
4	to	554	a	642	adoption	18	smuggler	7	teen	16	smuggler	7	trek	7
5	and	506	one	69	teen	16	songwriter	5	one	69	songwriter	5	teen	16
6	of	463	child	70	water	46	trek	7	family	56	trek	7	lawsuit	3
7	she	342	he	324	child	70	greed	10	adoption	18	greed	10	lure	13
8	in	341	who	101	desert	16	adoption	18	child	70	adoption	18	fertilization	3
9	he	324	that	304	corporate	17	allergy	8	lure	13	allergy	8	allergy	8
10	they	307	family	56	parent	28	distributor	13	water	46	distributor	13	greed	10
11	that	304	embryo	40	song	18	the	1060	distributor	13	mansion	10	courtroom	4
12	have	297	have	297	lure	13	say	242	vacuum	12	migrant	10	mitochondrial	2
13	I	254	water	46	say	242	be	800	greed	10	vacuum	12	scribble	2
14	say	242	find	53	distributor	13	a	642	syndrome	12	syndrome	12	flirt	2
15	for	182	do	148	letter	27	mansion	10	desert	16	donate	10	relay	3
16	it	171	tell	45	flight	16	she	342	they	307	salesman	8	asymmetrical	2
17	not	163	program	33	program	33	migrant	10	mansion	10	attendant	9	mansion	10
18	with	158	woman	39	airline	13	vacuum	12	migrant	10	say	242	rupture	2
19	do	148	out	58	vacuum	12	syndrome	12	corporate	17	sticker	6	purify	2
20	on	138	parent	28	sex	17	to	554	airline	13	airline	13	migrant	10
出現度数中央値		306		70		18		13		17		10		5
単語の平均文字数		2.6		4.0		5.8		5.5		5.9		6.8		7.4

表 5　指標別 *Japan Times* の特徴語上位 20 語

順位	頻度		補完類似度		ダイス係数		コサイン		対数尤度比		カイ二乗値 イエーツの補正公式		自己相互情報量	
1	the	1369	the	1369	firework	64	firework	64	firework	64	firework	64	firework	64
2	be	831	firework	64	marble	33	marble	33	marble	33	marble	33	frieze	6
3	of	620	one	61	percent	35	yen	18	percent	35	yen	18	yen	18
4	to	529	that	302	museum	34	postal	19	museum	34	postal	19	gunpowder	4
5	a	513	be	831	festival	26	percent	35	yen	18	percent	35	postal	19
6	and	466	government	50	fiscal	20	frieze	6	postal	19	frieze	6	pulp	5
7	in	410	year	70	postal	19	fiscal	20	festival	26	fiscal	20	marble	33
8	that	302	percent	35	yen	18	the	1369	fiscal	20	festival	26	depositor	4
9	have	266	say	104	movie	18	museum	34	one	61	museum	34	outlay	5
10	it	247	museum	34	island	22	festival	26	movie	18	pulp	5	sprout	2
11	I	234	marble	33	one	61	be	831	island	22	gunpowder	4	uninhabited	2
12	he	202	festival	26	democracy	17	of	620	sculpture	14	movie	18	funky	2
13	they	193	will	129	ministry	18	pulp	5	democracy	17	sculpture	14	fiscal	20
14	for	178	island	22	summer	22	gunpowder	4	ministry	18	outlay	5	archipelago	2
15	on	148	fiscal	20	sculpture	14	movie	18	summer	22	terrorism	8	specialty	2
16	with	147	summer	22	reform	18	to	529	government	50	depositor	4	priceless	3
17	will	129	postal	19	network	17	sculpture	14	frieze	6	democracy	17	propeller	2
18	as	121	cut	24	government	50	a	513	reform	18	antiquity	6	percent	35
19	you	114	people	45	rock	16	outlay	5	terrorism	8	airplane	7	restitution	2
20	say	104	around	28	stage	23	terrorism	8	network	17	island	22	terrorism	8
出現度数中央値		241		40		21		23		20		18		5
単語の平均文字数		2.7		5.4		6.4		5.4		6.7		7.0		7.5

表 4, **5** を一瞥すると，出現度数の中央値は，頻度，補完類似度，ダイス係数，以下，自己相互情報量まで，ABC News では 306 → 70 → 18 → 13 → 17 → 10 → 5 とほぼ段階的に少なくなり，*Japan Times* でも 241 → 40 → 21 → 23 → 20 → 18 → 5 とほぼ同様の傾向になる。一方，単語の文字数で見た長さの平均は ABC News では 2.6 → 4.0 → 5.8 → 5.5 → 5.9 → 6.8 → 7.4 と指標が左から右に移るにつれてほぼ段階的に長くなる。*Japan Times* でも 2.7 → 5.4 → 6.4 → 5.4 → 6.7 → 7.0 → 7.5 とほぼ同様の傾向が見られる。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ことから (竹蓋他,1994), 指標が左から右へ移るに従って，難易度の低い語から高い語へと段階的に特徴語が抽出されているのではないかということが推測できる。[13]

3.3 上位 300 語を対象にした平均文字数の累計的変化

前節で行なった上位 20 語の観察により,各指標の抽出単語には一定の傾向があることがわかった。本稿で調査している時事英語の資料規模と各指標の特徴が顕著に現れるかという点を考慮すると，資料に出現する語の種類の 1 割にあたる上位 300 語程度の各指標の抽出傾向を明確にしたいと考える。そこで前節で観察した単語の平均文字数を上位 300 語に拡大して考察を進めた。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 位からの文字数平均を 300

位まで累計的に求め，ABC News，*Japan Times*，それぞれの結果を**図 1** と**図 2** に示した。カイ二乗値とイエーツの補正公式はよく似た折線のため，カイ二乗値で代表した。その結果，折線は 7 本になっている。

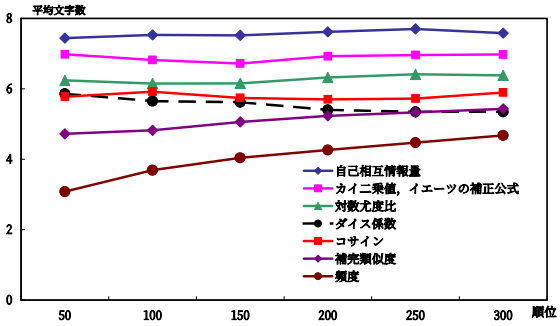

図 1　平均文字数の推移 (ABC News)

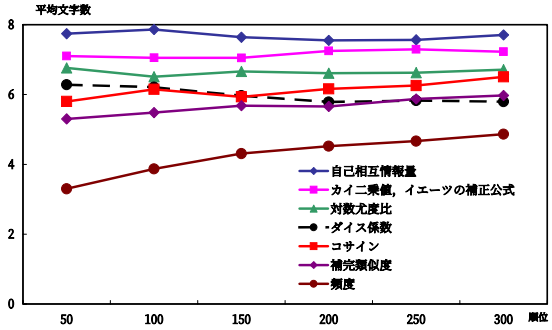

図 2　平均文字数の推移 (*Japan Times*)

図 1 と**図 2** から，前節の観察結果とほぼ一致して，上位 300 位までの場合にも 8 種類の指標はほぼ段階別に特徴語を抽出していることが観察できる。点線で示したダイス係数を除き，**表 4** と**表 5** の観察と同様に，自己相互情報量，カイ二乗値，イエーツの補正公式，対数尤度比，コサイン，補完類似度，頻度の指標順に長い語を抽出している傾向が確認できた。[14] これらの統計的指標は，語の長さの情報を指標値の算出に使用していないにもかかわらず，語彙レベルとの関連を示唆する結果が得られたことは興味深い。

3.4　特徴語の語彙レベルの比較

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 300 語の語彙レベルを比較する。

3.4.1　BNC 高頻度語の分布

各指標が一般分野の英語コーパスのどのような頻度帯(frequency band)に属する単語を抽出しているかを調査するため，BNC リスト 13,994 語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，**表 6** には ABC News の特徴語 300 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，**表 7** には *Japan Times* の場合を示した。300 語の 10%にあたる 30 語以上の特徴語が属する頻度帯を濃いグレーで，5-9%の特徴語が属する頻度帯を薄いグレーで示した。

結果，ABC News，*Japan Times* とも似たパターンを示した。両資料を通じ各指標の特徴語が多く属する頻度帯を見ていくと，頻度の指標は BNC1,000 語レベル，ダイス係数，補完類似度はほぼ BNC1,000～2,000 語レベルである。一方，対数尤度比，イエーツの補正公式の特徴語は BNC1,000～6,000 語レベルまで比較的均等に属している。コサインは 8,000～12,000 語レベルを除く頻度帯から，カイ二乗値は 8,000～10,000 語レベルを除く頻度帯から広く特徴語を選んでいる。自己相互情報量の特徴語の大部分は，ABC News では BNC4,000 語レベル以降に，*Japan Times* では BNC7,000 語レベル以降に属する。

表 6　BNC 頻度帯の分布(ABC News)

単位 %

BNC 頻度帯	頻度	ダイス係数	補完類似度	コサイン	対数尤度比	イエーツ補正	カイ二乗値	自己相互情報量
1,000	87	61	58	39	27	22	15	0
2,000	7	14	15	8	13	12	9	2
3,000	1	6	8	6	10	10	7	2
4,000	1	6	6	6	10	10	7	5
5,000	2	7	7	8	9	9	9	9
6,000	1	3	3	6	11	13	11	13
7,000	1	2	2	5	6	6	6	6
8,000	0	0	0	2	2	2	2	13
9,000	0	0	0	3	3	3	3	13
10,000	0	0	0	2	2	2	3	10
11,000	0	1	1	2	2	2	8	8
12,000	0	0	0	1	1	1	5	5
13,000	0	0	0	8	2	3	9	9
14,000	0	0	0	5	2	5	5	5

表 7　BNC 頻度帯の分布(*Japan Times*)

単位 %

BNC 頻度帯	頻度	ダイス係数	補完類似度	コサイン	対数尤度比	イエーツ補正	カイ二乗値	自己相互情報量
1,000	88	57	52	28	20	17	8	0
2,000	7	19	20	8	15	14	8	0
3,000	3	8	11	7	13	11	7	2
4,000	1	6	6	6	11	11	6	2
5,000	0	4	4	7	8	8	7	4
6,000	1	2	2	4	10	10	5	4
7,000	0	1	1	7	9	9	9	7
8,000	0	1	1	4	4	4	4	4
9,000	0	0	0	2	2	2	2	10
10,000	0	0	0	4	4	4	4	18
11,000	0	1	1	2	2	2	5	16
12,000	0	0	0	2	0	0	13	13
13,000	0	0	0	14	1	2	14	14
14,000	0	0	0	7	1	7	7	7

3.4.2　特徴語の学年分布

上位 300 語の特徴語を容易に理解できるのは，米国人の場合，どの学年くらいなのかを Dale & O'Rourke(1981)の資料にもとづいて調査した。ABC News についての結果を**図 3** に，*Japan Times* を**図 4** に示した。棒グラフは 300 語のうち何%の語が 4 年，6 年，8 年，10 年，12 年，13 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。

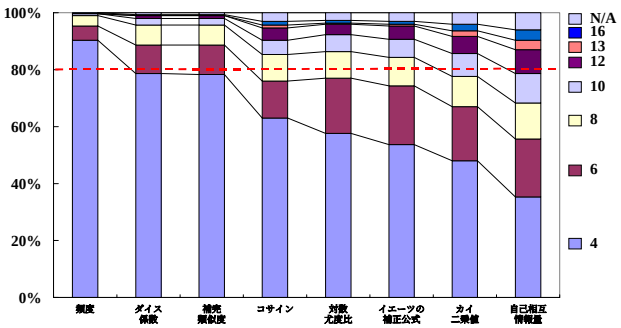

図 3 特徴語の学年分布(ABC News)

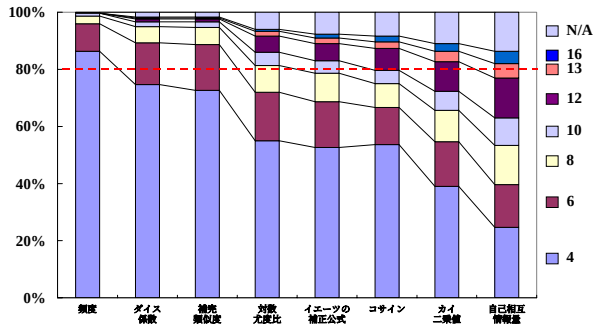

図 4　特徴語の学年分布(*Japan Times*)

仮に特徴語の 8 割が理解される学年を基準にして見ると，ABC News では，頻度によって抽出された特徴語は 4 年生，ダイス係数と補完類似度は 6 年生，コサイン，対数尤度比，イエーツの補正公式は 8 年生，カイ二乗値は 10 年生，自己相互情報量は 12 年生で理解されることになる。*Japan Times* では，頻度による特徴語は 4 年生,ダイス係数と補完類似度は 6 年生,対数尤度比は 8 年生，イエーツの補正公式は 10 年生，コサインとカイ二乗値は 12 年生，自己相互情報量の指標では 13 年生である。参考資料に含まれていない特徴語も多少あるが，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，統計的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。

3.4.3 学校英語教科書語彙との差

冒頭で述べたように，わが国の教育現場において時事英語を教材に使用する際には語彙の不足が常に問題となる。そこで，各指標の特徴語上位 300 語と中学校･高等学校教科書語彙との関連を検討する必要がある。ABC News と *Japan Times* の特徴語において中学・高校教科書の学習語彙には出現しない語の割合を算出した結果を**図 5**, **図 6** に示した。各指標が抽出した特徴語 300 語のうち，高校 3 年生までの教科書語彙で未習となる語の割合は，ABC News の場合，低い方から順に,頻度による抽出結果の 7%,ダイス係数 26%,補完類似度 27%,コサイン 46%,対数尤度比 52%,イエーツの補正公式 56%,カイ二乗値 67%,自己相互情報量が 85%であった。*Japan Times* の場合も未習となる語の割合がどの指標においても数％多いことと，コサインと対数尤度比の順序が入れ替わる以外は同様の傾向であった。この結果からも，各指標は明白に異なる学習段階レベルの特徴語を抽出していることがわかる。

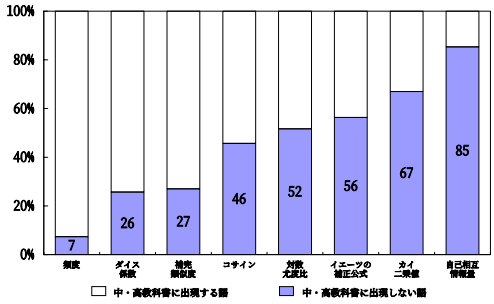

図 5　中・高教科書未習語(ABC News)

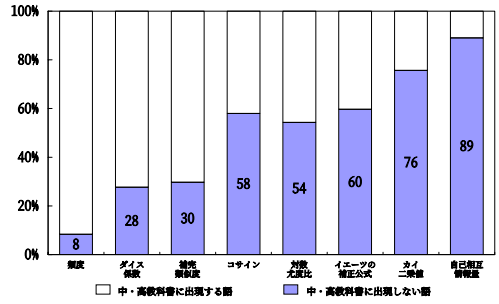

図 6　中・高教科書未習語(*Japan Times*)

図 5，**6** より得られた中・高教科書未習語の情報と，**3.1** 以降の考察からこれまで明らかになった情報を総合すると，各統計的指標を仮に英語習熟度に応じて段階的な学習用語彙選定に利用するとすれば，以下のようになる。[15]

①頻度による抽出結果では約 1 割が未習語，すなわち，9 割が既に学習した語であり，しかも大部分が BNC1,000 語レベルであるので，時事英語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入門レベル学習者に適していると考えられる。

②補完類似度とダイス係数による特徴語は BNC2,000 語レベルまでの語が中心であり，母語話者の 6 年生で特徴語の約 9 割が理解される。3 割近くある未習語も BNC 4,000～5,000 語レベルまでに属する基本的な当該分野の特徴語と考えられるので，全体として時事英語の初級レベル学習者向けと判断できる。

③対数尤度比，コサイン，イエーツの補正公式，カイ二乗値では BNC7,000 語レベルまで特徴語が分布しており，一部は 11,000 語レベルを超えるものもあること，さらに，未習語が 5 割前後を占めるので学習負荷が大きいことから，[16] 初級レベルを終えた中級レベル学習者対象の特徴語に適している。

④自己相互情報量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，そして抽出の中心は *Japan Times* の特徴語に顕著なように 7,000 語レベル以降の低頻度語であることから，時事英語の本格的習得を目指す上級レベル学習者に有効であると考えられる。

3.5 ABC News 1 単位ニュースにおける特徴語の実例

前節まで，ABC News，*Japan Times* の特徴語を全体的に観察した結果，各指標はそれぞれ特定の語彙レベルの特徴語を抽出していることが明らかになった。本節では，見方を変えて，1 単位ニュースの中で，実際に各統計的指標がどのような語を特徴語として選別しているかを見ていくことにする。

ここでは 2 万語の ABC News に含まれる 20 編のニュースからランダムに取り出した “A Tale of Two Families – Embryo Adoption Brings Them Together” とい

う 810 語(延べ語数)の長さのニュースをとりあげる。限られた紙面で傾向を観察しやすくするため，このニュースで使用された 281 語(異語数)から中学学習語彙を除き，次に検討対象を 8 種類の指標のいずれかが上位 300 位以内にランクした語に限定した。上位 300 位までに入らない語を除いた理由は，300 位より下位の語には日常的に必要度は高くても当該ニュースとは連想度の低い accept, ahead, amaze, beach, criminal 等が含まれるからである。また，中学学習語彙を除いた理由は，時事英語素材を学習する際には前提として中学学習語彙は既習語と考えられるからである。最終的に残った 35 語について，指標別の特徴語選別状況を**表 8** に「+」の記号で示した。**表 8** の特徴語は，抽出傾向を観察しやすいように，BNC リストを参照して各語に対応する BNC 頻度順位に従って配列した。

表 8　ABC News の 1 単位ニュースにおける特徴語上位 300 語の例

BNC頻度順位	いずれかの統計指標で300位より上位に抽出された語	頻度	ダイス係数	補完類似度	コサイン	対数尤度比	イエーツの補正公式	カイ二乗値	自己相互情報量
123	point	+							
129	follow	+							
349	research	+	+	+					
496	sign	+	+	+	+	+	+	+	
644	check		+	+					
657	couple	+	+	+	+	+		+	
758	legal		+	+		+			
777	treatment	+	+	+	+	+	+	+	
804	treat		+	+					
863	baby	+	+	+		+	+		
956	drug		+	+					
1048	option	+	+	+	+	+	+	+	
1146	accident	+	+	+	+	+	+	+	
1389	attract		+	+					
2225	awareness					+	+		
2473	twin			+		+	+		
2750	undergo		+	+	+	+	+	+	
2917	clinic					+			
2925	pregnant		+	+	+	+	+	+	+
3115	genetic		+	+	+	+	+	+	+
3126	biological					+	+		
3735	adoption	+	+	+	+	+	+	+	+
3988	donation		+	+	+	+	+	+	+
4334	fertility		+	+	+	+	+	+	+
4814	embryo	+	+	+	+	+	+	+	+
5061	nephew						+		+
5143	donate	+	+	+	+	+	+	+	+
5615	emotionally					+	+	+	+
7675	reunion								+
8551	earmark								+
8854	yearly				+	+	+	+	+
9016	enchant								+
9618	vitro				+	+	+	+	+
12129	adoptive				+	+	+	+	+
12807	fertilization				+	+	+	+	+

8 種の指標すべてに共通して選出された語は adoption, donate, embryo の 3 語であり，これらの語は ABC News のこの 1 単位ニュースのキーワードと言えよ

う。**表 8** における特徴語抽出の傾向は，**表 6** で観察した各指標の語彙レベル別の抽出パターンとほぼ符合しているようである。抽出された実例を**表 8** の左から右の指標へと眺めていく。頻度の指標で抽出された高頻度語は上述の共通 3 語を除いては，point, follow, research, baby などの一般的な語である。ダイス係数と補完類似度でも易しい汎用的な語が目に付く一方，pregnant，genetic，donation，fertility 等の当該ニュースをもっと具体的に説明するために必要な語が徐々に加わる。コサイン，対数尤度比，イエーツの補正公式，カイ二乗値では，research, check, drug 等の一般的なものが消え，vitro, adoptive, fertilization 等の当該ニュースを詳細に具体的に表現する語が抽出されている。表中の一番右の欄の自己相互情報量では，他の 7 指標で選ばれていた高頻度の汎用語は姿を消し，pregnant，genetic 以下に示される専門的な中～低頻度を中心とした語彙に厳選された抽出となっている。[17] 以上の具体的な単一ニュースの検討においても，8 種の統計的指標はそれぞれに固有の語彙レベル別に特徴のある語彙を抽出していること，また，特徴語の内容について，一般的なもの（general）からより具体的なもの（specific）へという抽出傾向が観察できた。以上，複数の統計的指標を適用して特徴語を抽出する方法は，学習者の英語習熟度や学習目的に応じた語彙選定の方法として高い実用性を有すると考えられる。

4. おわりに

本研究は,コーパス言語学や自然言語処理における最近の研究の発展を受け,統計的指標による特徴語抽出を，時事英語資料の語彙選定への応用という実用的観点から評価したものである。

本稿では，8 種類の統計的指標を利用して，比較的小規模な各 2 万語からなる 18 種の時事英語資料の特徴語を抽出し検討した。まず, 全体の指標間の順位相関を調査し，次に ABC News と *Japan Times* の各指標による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，特徴語の属する 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未習語の割合を調査した後， ABC News の 1 単位ニュースをとりあげて，具体的に特徴語の抽出状況の比較，検討を行なった。その結果，各統計的指標は 2 万語という比較的小規模の資料においても，また，1 単位ニュースにおいても，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。

実際に各統計的指標を学習者の英語習熟度に応じて段階的な学習用語彙選定に適用するとすれば，頻度は入門レベル，補完類似度とダイス係数は初級レベル，対数尤度比，コサイン，イエーツの補正公式，カイ二乗値は中級レベル，そして自己相互情報量は上級レベル学習者に応じた語彙選定に利用できると考える。従って，語彙の選定者が学習対象者の習熟度を考慮し，これらの統計的指標を上手に使い分ければ，効率的に時事英語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，時事英語の専門的知識を有しない英語教師にとっても，時事英語資料の特徴語を選定する際の一助になると期待できる。

注

＊　本稿をまとめるにあたり，千葉大学の西垣知佳子氏と日本大学大学院生の山﨑淳史氏にご協力いただきました。また，査読者の方々より有益なコメントをいただきました。心より感謝いたします。

1. 本稿での「語彙レベル」は，例えば BNC 頻度リストの 2,000 語レベル，3,000 語レベルのような客観的な目安を想定して用いている。語彙レベルは例えば Chujo(2004)で用いた方法により BNC を基準にして客観的に計測できる。
2. すべての未習語を補充するのは不可能なので，学習者の習熟度に応じて「指導，学習するとよい」語を選別し，残りの未習語は注釈(glossing)で対処する等が考えられる。
3. Scott, M. (1996) Wordsmith Tools, Ver 4. http://www.lexically.net/wordsmith/
4. http://www.corpora.jp/
5. 本稿での語彙リストの見出し語化の段階は分類するとすれば，齊藤他（1998:110-113）の「基底形の頻度表」と考えることができる。なお，固有名詞や数字は特定のテキストに集中して出現することが多いので（Nation, 2001:19-20），語彙リストの比較の際には除去されることが多い。これらの語は特徴度が非常に高いため，我々が直観的に考える特徴語の抽出には障害となる。本研究では，意味のある普遍的な結果を得るため，これらの語をすべての語彙リストから人手で取り除いた。ただし，実際の指導には，背景的

知識の説明とともに固有名詞の指導は必要と考える。

6. ここでの「時事英語資料」と「一般分野の英文資料」の区分の考え方は，齊藤他（1998）の「特殊目的コーパス」と「汎用コーパス」の区分に基づいている。「その研究目的によってサンプル・コーパスは，大きく２つに分けられる。すなわち，特定の言語研究のために編纂される特殊目的コーパス(special purpose corpus)と，さまざまな研究を想定して，いわば総合目的のために編纂される汎用コーパス(general purpose corpus)である」（齊藤他，1998：19-20）。なお，「時事英語資料」のような特殊目的コーパスに使われている特徴語は specialized vocabulary あるいは technical vocabulary と呼ばれ，その選定方法の 1 つとして本研究で使用した方法が Nation(2001)に示唆されている。“[O]ne way of making a technical vocabulary is to compare the frequency of words in a specialized text with their frequency in a general corpus” (Nation,2001:18)。
7. 自己相互情報量と対数尤度比の定義式では自然対数を使用している。本稿では，各指標値は特徴語の順位付けにのみ使用されているので対数の底は結論に影響しない。
8. 各統計的指標の工学的な説明を本稿に加えることが望ましいが，紙数の関係で到底不可能であるため，内山他(2004) (http://www2.nict.go.jp/jt/a132/members/mutiyama/pdf/chara.pdf)を参照されたい。
9. 指標値を求める時，自己相互情報量が BNC リストでの頻度ゼロの語を過大評価してしまうのを排除するため，時事英語リストの語彙から BNC13,995 語に含まれない語を除外した。それらの低頻度語は効率的な語彙学習の観点からも必要性は低いと考えられる。
10. 相加平均で順位相関係数の平均値がとれる理由は，18 種の資料の各々を時事英語からの１つのサンプルと考えることによって理解できる。各資料（サンプル）から１つの順位相関係数を計算し，それらの平均を求めることは，母集団における順位相関をサンプルの順位相関から推定していることになる。
11. 18 種の時事英語について，各指標により抽出されたすべての特徴語リストを掲載できないので **Appendix 2** に 18 種の特徴語リストの一部を付した。

12. イエーツの補正公式は低頻度語に対するカイ二乗値を補正するものであるため，本来，両者の性質は似ている。
13. **表 4**，**5** の出現度数と平均文字数の関係には「機能語」「内容語」という側面からの観察も重要と思われる。しかし，本研究の目的である「特定分野の英語に特徴的な語彙の選定」は実質的には内容語を対象とするものであるので，非特徴語である機能語についてここでは言及していない。
14. **図 1**，**2** においてダイス係数が他とは異なり，順位が下がるにつれて平均文字数が低下している原因は今後詳しく検討したい。
15. 本稿で目指している統計的指標を利用した語彙選定の位置づけについて我々は次の 2 段階の方法を想定している。 1)特定分野における単語をその特徴度によって順位付ける。 2)そ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（post-filtering)。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)の語彙選定の第一次資料(primary resources)を精度良く機械的に得ることを目指すものである。①～④の考察は 1)の段階についてである。
16. 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。実際には時事英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。
17. 自己相互情報量で抽出された語は,他の指標の特徴語と比較して相対的に専門的な語彙と考えられる。しかしながら自己相互情報量は低頻度語間の関連度を過大評価することが知られている(Manning et al.,1999)。比較的専門的な語彙を抽出する一方,一般的な語彙の低頻度語も特徴語として抽出される。

参考文献

東眞須美(1998)「大学における時事英語を使った授業」『英語教育』47, 8: 17-19.

中條清美・長谷川修治(2003)「時事英語の授業で用いられる英文素材の語彙レベル調査」『時事英語学研究』42: 51-62.

Chujo, K. (2004) “Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List.” In J. Nakamura, N.

Inoue, & T. Tabata (eds.) *English Corpora under Japanese Eye*. Amsterdam: Rodopi, pp.231-249.

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.

Dale, E. & J. O'Rourke (1981) *The Living Word Vocabulary*. Chicago: World Book-Childcraft International, Inc.

Dunning, T. E. (1993) "Accurate Methods for the Statistics of Surprise and Coincidence." *Computational Linguistics*, 19,11: 61-74.

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』東京：JACET.

Hisamitsu, T. & Y. Niwa (2001) "Topic-Word Selection Based on Combinatorial Probability." *NLPRS-2001*: 289-296.

市川泰男・安吉逸季・J. R. Hestand・塩川春彦・小林千春・萩野敏 (2003)『*Unicorn English Course Ⅰ,Ⅱ*』東京：文英堂.

市川泰男・安吉逸季・J. R. Hestand・塩川春彦・小林千春・萩野敏 (2003)『*Unicorn English Reading*』東京：文英堂.

笠島準一他 (2002)『*New Horizon English Course 1, 2, 3*』東京：東京書籍.

Manning, C. D. & H. Schutze (1999) *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

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

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

Rixon, S. (1998) "Authenticity." In Johnson K.et al. (eds.), *Encyclopedic Dictionary of Applied Linguistics*. Oxford : Blackwell Publishers Ltd, pp.68-69.

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』東京：研究社出版.

谷口幸夫(1998)「高校でできる時事英語を使った授業」『英語教育』47, 8: 11-13.

竹蓋幸生(1981)『コンピューターの見た現代英語：ボキャブラリーの科学』東京：エデュカ.

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.

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.

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.

Appendix 1　使用した統計的指標の定義式

対数尤度比：$LLR_0 = a\log(an/((a+b)(a+c))) + b\log(bn/((a+b)(b+d))) + c\log(cn/((c+d)(a+c))) + d\log(dn/((c+d)(b+d)))$

カイ二乗値：$Chi2_0 = \left(n(ad-bc)^2\right)/((a+b)(c+d)(a+c)(b+d))$

イエーツの補正公式：$Yates_0 = n(|ad-bc|-n/2)^2/((a+b)(c+d)(a+c)(b+d))$

上記３指標の補正：$LLR = sign(ad-bc)\times LLR_0$

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

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

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

頻度：$Freq = a$

ダイス係数：$2a/((a+b)+(a+c))$

補完類似度：$CSM = (ad-bc)/\sqrt{(a+c)(b+d)}$

コサイン：$Cosine = a/\sqrt{(a+b)(a+c)}$

自己相互情報量：$MI = \log(an/((a+b)(a+c)))$

Appendix 2　18 種の時事英語資料の自己相互情報量による特徴語(上位 20 語)

順位	ABC News	CBS News	CNN News	PBS News	BBC News	VOA News	VOA Special English	やさしいビジネス英語	*Time*
1	songwriter	railroad	rescuer	turret	lade	peacekeeping	migraine	teen	abduct
2	smuggler	haircut	glacier	retaliate	racer	tenet	mister	healthcare	laity
3	embryo	thaw	sadden	oversight	carnival	terrorism	uterus	web	homeland
4	trek	estrogen	oversight	tariff	obesity	grenade	steroid	online	testify
5	teen	awsuit	parole	mosquito	seeker	televise	adenoma	acupuncture	aide
6	lawsuit	soy	pediatric	incumbent	terrorism	topple	tuberculosis	high-tech	fingerprint
7	lure	adversarial	stoppage	dent	firefighter	oust	malaria	gourmet	diocesan
8	fertilization	bombing	anti-aircraft	booster	peacekeeping	landless	iceberg	clout	bandit
9	allergy	fluke	southwest	state-of-the-art	endanger	gunman	genome	fickle	pastor
10	greed	menopause	kidnap	escalation	commemoration	sabotage	mosquito	well-informed	gag
11	courtroom	zip	sheriff	collateral	factional	eviction	inactive	info	forgo
12	mitochondrial	hormone	assassinate	technologically	separatist	bloodshed	shuttle	closet	governance
13	scribble	herbal	abduction	columnist	defection	consulate	lizard	yoga	flirt
14	flirt	air-conditioning	recount	governance	arsenal	aggressor	infect	high-speed	deluge
15	relay	unopened	abduct	inferno	spacecraft	indict	perspiration	gadget	indict
16	asymmetrical	barber	canine	forestall	pipeline	dick	transplant	burner	kidnap
17	mansion	abbreviation	gunman	cedar	smuggler	reformist	headache	surf	diocese
18	rupture	undetected	miller	jeopardize	heartfelt	sniper	bacterium	trash	ordain
19	purify	gadget	indict	citrus	obese	evaporate	hurricane	enroll	prod
20	migrant	inventor	perpetrator	trade-off	sanitation	semiconductor	platelet	perk	sprawl

順位	*Newsweek*	*New York Times*	*USA TODAY*	*Chicago Tribune*	*Japan Times*	*Daily Yomiuri*	*Asahi Weekly*	*Student Times*	*News For You*
1	fugitive	cent	mosquito	lawsuit	firework	yen	yen	lice	teen
2	ranch	inauguration	tuition	cent	frieze	quilt	curd	soccer	lettuce
3	baseball	hormone	salmonella	juror	yen	governance	soy	euthanasia	railroad
4	container	assassinate	lawsuit	ranch	gunpowder	dispenser	bureaucrat	yen	diner
5	disarm	homemade	barricade	kidnap	postal	indict	issuance	soy	baseball
6	heartland	year-end	voucher	prosecutor	pulp	privatize	hands-on	hockey	ford
7	toll	improperly	enrollment	northwest	marble	slowdown	teller	subway	shooting
8	longtime	obesity	abatement	funnel	depositor	inflow	secretariat	infestation	athlete
9	briefing	eviction	rescuer	testify	outlay	estrogen	abduct	basketball	aerobic
10	smuggler	sham	leftist	percent	sprout	modernize	statue	ranch	robot
11	aide	congressional	pediatric	marijuana	uninhabited	candidacy	spooky	cuisine	pearl
12	web	obese	inauguration	leverage	funky	residency	urchin	maid	lawsuit
13	handcuff	grieve	creek	online	fiscal	fraudulent	referee	jubilee	handcuff
14	fort	leftist	passport	yen	archipelago	upstream	abduction	baseball	gunman
15	martyr	gulf	virus	gag	specialty	tariff	bout	thug	hawk
16	reprisal	inaugurate	non-profit	airline	priceless	procure	displacement	shampoo	terror
17	mosque	slum	trash	wireless	propeller	corp	unveil	elevator	kidnapper
18	atop	exam	oversight	attorney	percent	dam	assailant	craze	juror
19	truck	ranch	enroll	innermost	restitution	getaway	automobile	spicy	poppy
20	killing	categorize	deceptive	undervalued	terrorism	malpractice	emperor	ventilator	kidnap

中條清美（日本大学　chujo@cit.nihon-u.ac.jp）
内山将夫（情報通信研究機構　mutiyama@nict.go.jp）
長谷川修治（千葉県立長狭高等学校　shase@alto.ocn.ne.jp）