

日本語教育語彙を選定するための統計的指標 — 尤度比検定、カイ2乗検定、イエーツの補正公式の特徴 —

寺嶋 弘道

アブストラクト：

国立国語研究所による「現代日本語書き言葉均衡コーパス モニター公開データ（2008年度版）」が利用できるようになり、日本語教育分野でも大規模コーパスを利用し、日本語教育語彙を選定しようとする動きが始まっている。英語教育分野では、早くからそのような動きが見られ、多くの研究者が関わった「JACET8000」はその代表的なものである。「JACET8000」では、統計的指標を用いて、コーパスの中で特徴的に使用されている語を明らかにし、それを語彙選定の目安にする方法が取り入れられている。今後、日本語教育分野でも大規模コーパスを用いて、語彙選定を行う動きが活発になると考えられるため、本稿では、尤度比検定、カイ2乗検定、イエーツの補正公式という3つの類似した統計的指標の特徴を考察した。

その結果、尤度比検定は統計量を抑える指標であること、イエーツの補正公式は統計量が安定していない指標であること、カイ2乗検定は他の2指標に比べ、常に統計量が高くなる指標であることが確認できた。

キーワード：語彙選定、特徴語、尤度比検定、カイ2乗検定、イエーツの補正公式

1. はじめに

現在、国立国語研究所では「現代日本語書き言葉均衡コーパス（以下BCCWJ）」を構築しており、2011年には1億語規模の大規模コーパスが完成する予定である。2008年からは、著作権処理が済んだデータの一部が「BCCWJモニターデータ（2008年度版）」として利用できるようになり、日本語学、日本語教育学の分野でも大規模コーパスを応用しようという動きが始まっている。BCCWJの特徴は、書籍、白書、Yahoo知恵袋といった多様な内容が含まれていること、そして、母集団の特性が反映されるよう、構築されていることである。

丸山（2009）は日本語教育でのBCCWJの利用方法の一つとして語彙頻度表を挙げているが、その頻度表は日本語教育語彙の選定のための基礎資料として期待できるものである。たとえば、一般日本語（Japanese for General Purposes）用の語彙選定であれば、より多くのテキストに含まれる高頻度の語彙を抽出することで、統計的に裏付けされた選定が行える。一方、目的別日本語（Japanese for Specific Purposes）用の語彙選定においても特定目的のもとで構築されたコーパスを大規模コーパスであるBCCWJに参照すれば、統計的に偏りがある特徴的な語（以下特徴語）を明らかにできる。

英語教育分野ではそのような大規模コーパスを用いた語彙選定が既に行われている。多くの研究者が参加した大学英語教育学会基本語改訂委員会編（2003）による「JACET8000」では、コンピュータによる言語処理や統計的指標を駆使し、1億語のBritish National Corpus（以下BNC）、及び日本人の英語教育に必要なと考えられる言語資料等から日本人英語学習者のための教育語彙をレベル別に選定している。この選定では、BNCから得られた語彙頻度表を基準データとし、その基準データを英語教育の観点から集めたコーパスに参照させ、尤度比検定（以下LLR）という統計的指標でBNCから得られた順位を調整するという方法が取り入れられている。

LLRとは、2つのコーパスの中で統計的に強い偏りがある特徴語を抽出するためのもので、WordsmithやAntconcといったコーパス分析ツール¹でも用いられている統計的指標である。

日本語教育の分野では、橋本(2009)が「BCCWJモニターデータ(2008年度版)」によって特徴語の抽出を行っている。同研究では近藤(2009)でも用いられ、Kilgarriff(1996)が示したLLRの計算式に基づき、話題別特徴語の抽出を試みている。今後、BCCWJの完成に伴い、そのような統計指標を用いる動きは多くなっていくと考えられる。しかし、日本語教育の分野では、コーパスからの特徴語抽出の指標としてLLRが適切かどうか論じられたことがなく、同じ有意水準と臨界値を使用するカイ2乗検定(以下Chi2)及びイエーツの補正公式(以下Yates)との違いを報告したものもない。

そこで、本稿では今後、日本語教育語彙の選定を行うために必要になると考えられ、特徴語を抽出することができるLLR、Chi2、Yatesの3つの指標にどのような違いがあるのかを考察したい。具体的には、各指標に基づき抽出された特徴語の違い、統計量の高さの相対的關係、低頻度語における統計量の高さ、コーパスサイズの統計量への影響を調査し、3つの指標の特徴を明らかにしたい。なお、LLRは、大学英語教育学会基本語改訂委員会編(2003)で用いられたものもあるが、本稿ではKilgarriff(1996)で示されたものを対象とする。

2. 先行研究

あるコーパスに特徴的に現れる語を抽出するためには、まず、対象コーパスと参照コーパスを準備する必要がある。

対象コーパスは特徴語抽出の対象となるコーパスであるため、特定目的のもとで集められたコーパスを用いることが多い。それに対し、参照コーパスは参照するためのコーパスであるため、バランスがとれた大規模コーパスを用いることが多い。両コーパスから得られた語彙頻度表に基づき、表1のような2×2の分割表を作ると、下記に示した<計算式>で統計量を求めることができる。ここで求めた統計量の高さに応じ、特徴語が抽出される。

表1 統計量を求めるための分割表

	対象コーパス	参照コーパス	計
見出し語Wの頻度	a	b	a+b
見出し語W以外の頻度	c	d	c+d
計	a+c	b+d	a+b+c+d= (n)

c = 対象コーパスの総語数 - a

d = 参照コーパスの総語数 - b

<計算式>

① $Chi2 = n(ad-bc)^2 / ((a+b)(c+d)(a+c)(b+d))$

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

③ $LLR = 2(a \log(a) + b \log(b) + c \log(c) + d \log(d) - (a+b) \log(a+b) - (a+c) \log(a+c) - (b+d) \log(b+d) - (c+d) \log(c+d) + n \log(n))$

表1からもわかるように、統計量は2×2の分割表を用いた独立性の検定を応用したものである。このような独立性の検定で最もよく知られているものが①の計算式で求められるChi2である。しかし、統計学的に2×2の分割表の場合、「理論的期待度数が5以下の場合には特別な配慮が必要であ

る」(池田他1989, P99)という指摘や「 χ^2 を使って検定するのは好ましくない」(太郎丸2005, P46)という指摘もあり、②の計算式で補正を行うことが提案されている²。ただ、それらの指摘は2×2の分割表に1000万語や1億語規模のデータを用いるコーパス言語学的な立場を前提にしていなと思われる。

一方、コーパス言語学的な立場では、特徴語を抽出する場合、 χ^2 は統計量が高くなりすぎると Kilgarriff (1996) が指摘している。このような背景から、同研究では、Dunning (1993) が示した度数が1であっても正しい数値が得られるという LLR の計算式を③の形で紹介している³。国内の英語教育分野においては、中條他 (2004) が LLR、Chi2、Yates を含む 8 つの統計的指標、及びそれを組み合わせた総合指標による特徴語抽出に関する報告をしている。その結果、LLR、Chi2、Yates によって抽出された特徴語のリストは、順位相関係数が 0.91 以上で、非常に相関性の強い統計指標であるという。また、統計量の高い上位 20 語のみの考察ではあるが、LLR が TOEIC での特徴的な単語をよく抽出し、Chi2、Yates が LLR よりもレベルの高い特徴語を抽出していると報告している。しかし、同研究では池田他 (1989) を参考した LLR の計算式に 2 が乗じられていないという問題があると思われる⁴。以後、中條他 (2005) においても統計的指標に関する報告があるが、同様の問題が見られ、さらに詳しい調査の必要性が感じられる。

3. 研究方法

3. 1 研究目的

本稿では、特徴語を抽出することができる LLR、Chi2、Yates の 3 つの指標にどのような違いがあるのかを考察したい。そこで、実際に対象コーパスと参照コーパスから作成した語彙頻度表に基づき、各指標の統計量を求め、1) 各指標に基づき抽出された特徴語の違い、2) 統計量の高さの相対的關係、3) 低頻度語における統計量の高さ、4) コーパスサイズの統計量への影響という 4 点について調査を行い、3 指標の特徴を探りたい。以下、調査したい 4 点について詳しく述べる。

- 1) 統計量が高い語から順に特徴語として抽出した場合、抽出した語は各指標でどのぐらい違うのか。また、抽出範囲によってどのような違いが見られるか。
- 2) 3 指標の統計量の高さは相対的にどのような関係になるのか。また、統計量の高さによって相対的な関係に変化が見られるか。
- 3) 対象コーパスにおいて低頻度語である場合、統計量の高さは相対的にどのような関係になるのか。参照コーパスにおいて低頻度である場合と高頻度である場合、その相対的な関係に変化が見られるか。
- 4) 参照コーパスの総語数を 1 億語と換算し、対象コーパスの総語数を 10 万語、100 万語、1000 万語に換算した場合、3 指標の統計量の高さは相対的にどのような関係になるのか。その相対的な関係に変化が見られるか。

3. 2 対象データと言語処理

まず、本調査で使用した実験データについて述べる。本調査では、「BCCWJ モニターデータ (2008 年度版)」の中に含まれる書籍データと白書データを用いた。書籍データは、2001 年から 2005 年に発行された書籍文字を母集団としているもの、及び 1986 年から 2005 年に発行された書籍のうち、東京都内のより多くの公共図書館で共通に所蔵されている書籍に含まれる文字を母集団としているものがある。それぞれの書籍データは、母集団の量的構造が反映できるよう、比例配分によって抽出する文字数が決められ、無作為抽出されている。一方、白書データは、政府刊行物の白書のうち、1976 年から 2005 年までの 30 年間の全てを対象に無作為抽出されたものである。

書籍データと白書データのうち、どちらを参照コーパスにするかという問題があるが、含まれてい

るサンプル数の多さ及びその多様性を比べた場合、書籍データを参照コーパスとして考えることが適切だと思われる。そこで、本稿では書籍データを参照コーパスとし、対象コーパスである白書データに含まれる特徴語を抽出することにした⁵。以下、書籍データを参照コーパス、白書データを対象コーパスと呼ぶことにする。

次に、言語処理の方法について述べる。「BCCWJモニターデータ (2008年度版)」には、UniDicにより形態素解析されたデータが含まれている。本調査では、そのデータを処理することで、統計量を抽出するために必要な頻度情報を得た。その結果の一部を表2に示す。なお、頻度数の計算とマージに関する言語処理は、全てActive Perl5.10を用いた。

次に、見出し語が対象コーパスにどの程度偏りがあるのかを明らかにするため、対象コーパスの見出し語の頻度を参照コーパスのものと照らし合わせ、前述した計算方法で統計量を計算した。頻度情報から統計量を求める言語処理は、Microsoft Office Access 2003を用いた。ただし、両コーパスのうち、どちらへ偏りがあるのかを明確にするために、内山他 (2004) のように、表1の分割表において $ad-bc < 0$ の場合に①、②、③の計算式に-1を乗じる補正を行った。

表2 各コーパスの分析結果

	参照コーパス： 書籍データ	対象コーパス： 白書データ
サンプル数	4248	1500
総語数	11,964,405	4,881,858
異なり語数	73,473	27,762

* 総語数及び異なり語数は空白・記号を除く

4. 分析

4.1 特徴語の抽出と各指標の共通性

特徴語を抽出する場合、指標とその抽出範囲を決めることが必要である。指標ごとで抽出される特徴語が違うのは当然であるが、その抽出範囲も選定に影響を与える要因である。そこで、最初の調査では、統計量の高い特徴語から抽出した場合、抽出範囲によってどのような違いが現れるのかを調査した。

ここで問題となるのは、上位何位までのデータを対象にするかということである。まず、その基準を決めるにあたり、3つの指標の平均値を出し、その平均統計量の高さで並び変えることにした。後述する4.2の調査では、各指標の相関性の強さが視覚的にわかり、計算される平均統計量は各指標と相関性が高いと思われる。したがって、この調査で一時的に平均統計量を求めることは大きな問題がないと考えた。

次に、カイ2乗分布表で自由度1、5%で有意になる場合の臨界値を調べたところ、その数値は3.84であった。調査データで平均統計量が3.84になるのは、7720位であったため、1000単位でデータを抽出し、8000位までを抽出対象とした。1000単位であれば、3指標の統計量の違いが十分に現れると考えたためである。また、データ数に比例し、統計量が高くなる特徴語の抽出では、高見 (2003) で7.88 ($p < .005$)、近藤 (2009) で10.83 ($p < .001$) といったように非常に高い臨界値を使用することが多いことから、平均統計量が3.3である8000位までを抽出範囲とするのは、十分だと考える。ただし、統計量が同じ場合もあるため、その場合は同順位としてカウントするという方法を用いている。

表3は、各指標の統計量の高さによって決められた順位に基づき、特徴語を抽出した場合、指標間

でどれぐらいの語が共通していないかを数えたもので、抽出範囲によつての違いを示したものである。以下、本稿では指標間で共通していない語のことを非共通語と呼ぶことにする。

表3 各指標間での非共通語の数

	順位	LLR	Chi2	Yates
LLR	1-1000位まで	—		
	1-2000位まで			
	1-3000位まで			
	1-4000位まで			
	1-5000位まで			
	1-6000位まで			
	1-7000位まで			
	1-8000位まで			
Chi2	1-1000位まで	5	—	
	1-2000位まで	16		
	1-3000位まで	35		
	1-4000位まで	51		
	1-5000位まで	164		
	1-6000位まで	0		
	1-7000位まで	0		
	1-8000位まで	7		
Yates	1-1000位まで	5	1	—
	1-2000位まで	18	4	
	1-3000位まで	36	8	
	1-4000位まで	50	26	
	1-5000位まで	164	43	
	1-6000位まで	361	366	
	1-7000位まで	831	831	
	1-8000位まで	0	0	

まず、表3において各指標間の非共通語を合計した場合、LLRとChi2の指標間で最も非共通語が少なく、指標間での共通性が高いことがわかる。しかし、5000位までを範囲とした場合には非共通語が多く見られるため、それは抽出範囲によって異なるといえる。

一方、各指標間の非共通語を合計した場合、最も非共通語が多かったのはLLRとYatesの間であった。これは、2指標間の共通性が低いということになる。それらの指標間で非共通語が最も多くなるのは1位-7000位までで、次いで1位-6000位まで、1位-5000位までであった。また、1-8000位までを対象にすると、LLRとYatesの指標間では、非共通語が0になるという興味深い結果になった。この結果から、LLRとYatesの指標間では、下位で順位の違いが大きくなるが、その違いは、ある一定の範囲内で起きている現象だということになる。

4. 2 統計量の高低で見られる特徴

Kilgarriff (1996) では、Chi2の統計量の高さが指摘されている。そこで、次の調査では、3指標の統計量の高さが相対的にどのような関係になるのか、また、その統計量の高さによってその相対的な関係に変化が見られるかを調査した。

データは、4.1で用いたものを使用した。このデータは、1000単位で8つのデータに分けられているため、各群に乱数を発生させ、無作為に10語ずつ、計80語を抽出した。無作為抽出した80語の各頻度と各統計量を表4に示し、そのグラフを図1から図8に示す。横軸は表4での各語のID、縦軸は統計量を表している。以下では、統計量の相対的な関係を視覚的に確認するため、図を中心に分析

する。

まず、図1と図2から、Chi2とYatesがほぼ同程度の高さを示していることがわかる。LLRも類似した変化をしているものの、その2指標よりも統計量が低くなっている。図1と図2のケースでは、LLRの統計量は、他の2指標よりも最大で約12%低くなっている⁶。Chi2の統計量が高いという先行研究での指摘は、この結果からも確認できるが、Yatesも同様に高い統計量を示している。

図3と図4からは、全体の統計量が100以下になっている。これらの図からは徐々にYatesの統計量が低くなっており、YatesがLLRを下回ることも確認できる。Chi2は変わらず、3指標の中で最も高い統計量を示している。さらに統計量が低くなる図5から図8でYatesの統計量の低さが顕著になり、その高低の変化も他の2指標と異なっていることがわかる。また、LLRとChi2を比べると、統計量が低いケースにおいてもChi2のほうが高く、高低の変化もよく似ていることがわかる。

表4 無作為抽出した特徴語の統計量

	ID	見出し語	白書	書籍	LLR	Chi2	Yates
1-1000位まで	1	取り締まり	629	356	513.00	582.26	580.57
	2	産品	247	28	450.06	494.58	491.63
	3	以上	3817	4157	1261.33	1383.06	1382.14
	4	スポーツ	818	710	401.69	447.69	446.50
	5	共同	1568	926	1227.93	1392.15	1390.50
	6	価格	3005	1008	3611.13	4109.43	4107.19
	7	被災	305	75	429.37	485.62	483.13
	8	伸び	1919	117	3938.76	4215.50	4212.33
	9	死傷	311	50	514.23	573.32	570.54
	10	歩行	279	73	381.74	432.43	429.99
1001-2000位まで	1	離職	161	28	259.43	290.11	287.39
	2	共管	46	0	113.95	112.74	109.31
	3	稼働	150	104	99.02	111.64	110.18
	4	日数	165	80	153.96	175.24	173.38
	5	港	275	201	170.47	191.76	190.36
	6	ウエート	112	29	153.99	174.40	171.95
	7	即応	136	23	221.21	247.11	244.37
	8	養殖	196	69	228.83	260.55	258.37
	9	所定	211	45	315.45	355.27	352.68
	10	休業	101	41	107.57	122.57	120.53
2001-3000位まで	1	テレホン	33	5	55.58	61.82	59.04
	2	魚価	19	0	47.07	46.57	43.18
	3	音声	163	184	49.94	54.60	53.73
	4	滋賀	91	64	59.08	66.57	65.13
	5	子牛	41	8	63.43	71.22	68.59
	6	鋭意	39	4	72.73	79.59	76.62
	7	商用	38	7	60.02	67.27	64.60
	8	恐喝	61	25	64.54	73.54	71.52
	9	肉類	52	23	52.09	59.34	57.40
	10	転廃業	27	0	66.88	66.17	62.77

日本語教育語彙を選定するための統計的指標
 — 尤度比検定、カイ2乗検定、イエーツの補正公式の特徴 —

	ID	見出し語	白書	書籍	LLR	Chi2	Yates
3001-4000位まで	1	ブルガリア	26	9	30.66	34.91	32.74
	2	コモン	18	3	29.42	32.84	30.15
	3	賦課	35	22	25.73	29.12	27.56
	4	放火	94	102	31.28	34.31	33.39
	5	I N S A G	10	0	24.77	24.51	21.18
	6	十	25	12	23.52	26.77	24.93
	7	教誨	16	4	22.36	25.30	22.88
	8	一巡	24	12	21.84	24.84	23.05
	9	盗難	36	21	28.53	32.35	30.71
	10	養育	121	151	29.33	31.78	31.03
4001-5000位まで	1	喚起	58	71	14.75	16.01	15.24
	2	物損	12	2	19.61	21.90	19.23
	3	船位	9	0	22.29	22.06	18.74
	4	全域	46	57	11.35	12.31	11.56
	5	浸透	119	195	11.53	12.14	11.71
	6	転ずる	27	24	12.79	14.23	13.09
	7	田沢湖	9	2	13.23	14.92	12.47
	8	県外	23	16	15.12	17.05	15.62
	9	付則	28	18	20.10	22.73	21.21
	10	要領	131	208	14.56	15.38	14.92
5001-6000位まで	1	格付け	28	34	7.26	7.89	7.12
	2	過熱	20	17	10.13	11.30	10.12
	3	県土	4	0	9.91	9.80	6.66
	4	昇任	12	7	9.51	10.78	9.19
	5	マネタリー	4	0	9.91	9.80	6.66
	6	C R D	4	0	9.91	9.80	6.66
	7	各面	7	2	9.17	10.41	8.18
	8	裏作	9	4	8.98	10.23	8.37
	9	深度	26	26	10.11	11.16	10.17
	10	弁務	6	1	9.81	10.95	8.36
6001-7000位まで	1	N T I A	4	1	5.59	6.32	4.09
	2	受水	3	0	7.43	7.35	4.31
	3	漁場	4	1	5.59	6.32	4.09
	4	助産	9	7	5.16	5.78	4.53
	5	波力	3	0	7.43	7.35	4.31
	6	受け付ける	51	84	4.83	5.08	4.66
	7	自所	3	0	7.43	7.35	4.31
	8	海面	59	96	5.90	6.22	5.78
	9	P H Y	2	0	4.95	4.90	2.06
	10	精錬	9	7	5.16	5.78	4.53
7001-8000位まで	1	WH	2	0	4.95	4.90	2.06
	2	通則	3	1	3.62	4.12	2.18
	3	脱着	3	1	3.62	4.12	2.18
	4	那賀	3	1	3.62	4.12	2.18
	5	憂慮	22	32	3.40	3.63	3.08
	6	新增設	2	0	4.95	4.90	2.06
	7	I S D S	2	0	4.95	4.90	2.06
	8	歳計	3	1	3.62	4.12	2.18
	9	精留	2	0	4.95	4.90	2.06
	10	両港	2	0	4.95	4.90	2.06

図1 1-1000位までの無作為抽出

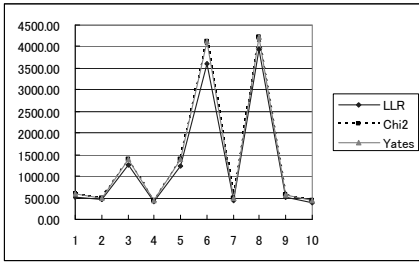


図2 1001-2000位までの無作為抽出

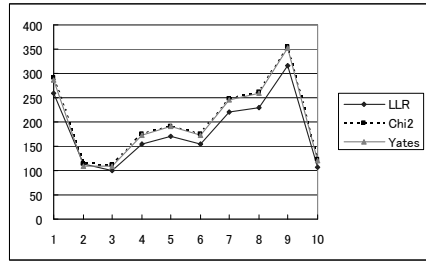


図3 2001-3000位までの無作為抽出

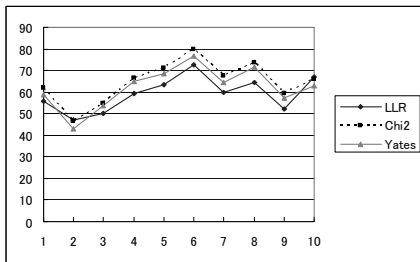


図4 3001-4000位までの無作為抽出

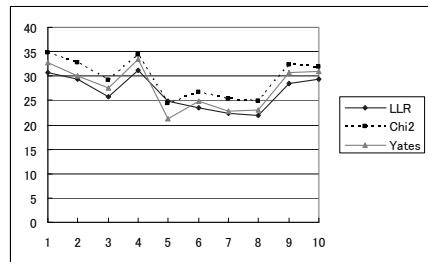


図5 4001-5000位までの無作為抽出

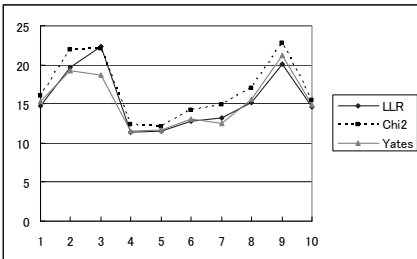


図6 5001-6000位までの無作為抽出

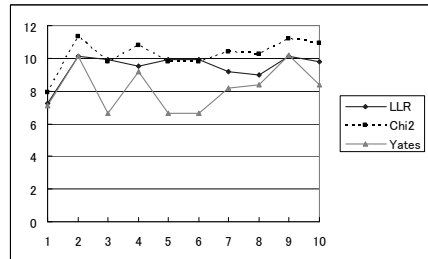


図7 6001-7000位までの無作為抽出

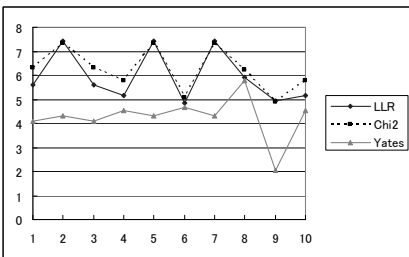
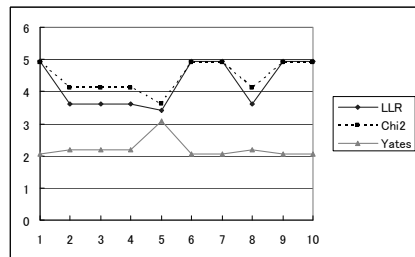


図8 7001-8000位までの無作為抽出



4. 3 低頻度語で見られる特徴

Dunning (1993) は、頻度が低い現象と高い現象とを直接的に比較できる指標としてLLRを提案した。特徴語の抽出では、対象コーパスで頻度が低い場合、その語の統計量が高くなりすぎないか注意が必要である。そこで、次の調査では、対象コーパスで見出し語が低頻度である場合、その統計量の高さが相対的にどのような関係になるのかを調査した。ここでは、参照コーパスが低頻度である場合と高頻度である場合では、異なる結果になると考えたため、参照コーパスが低頻度である場合と高頻度である場合に分け、調査を行った。低頻度と高頻度の場合、統計量はマイナスの数値になるが、頻度やレンジといった他の指標と組み合わせて特徴語を抽出する方法も考えられるため、ここで調査したい。なお、本稿では低頻度語を1以上5以下の語、高頻度語を100以上とする。

対象コーパスでの頻度が1以上5以下で、参照コーパスでの頻度も1以上5以下である群、そして、対象コーパスでの頻度が1以上5以下で、参照コーパスでの頻度も100以上である群のそれぞれに乱数を発生させ、無作為に10語ずつ、計20語を抽出した。無作為抽出した20語の各頻度と各統計量を表5に示し、そのグラフを図9と図10に示す。横軸は表5での各語のID、縦軸は統計量を表している。以下では、統計量の相対的な関係を視覚的に確認するため、図を中心に分析する。

まず、図9からわかるように、低頻度語同士での統計量は、 $Yates < LLR < Chi2$ となり、その高低の変化もよく似ていることがわかる。3指標の中では、Yatesの統計量の低さが目立つのに対し、LLRとChi2は統計量の高さが似ているといえる。一方、図10では、高低の変化は似ているものの、統計量が $LLR < Chi2, Yates$ となり、特にLLRの統計量の低さが目立っている。これらの結果から、低頻度同士で統計量を計算するケースでは、Yatesが最も統計量を抑える指標だが、低頻度と高頻度で統計量を計算するケースでは、LLRが最も統計量を抑える指標だといえる。

表5 無作為抽出した低頻度語の統計量

	ID	見出し語	白書	書籍	LLR	Chi2	Yates
低頻度と低頻度	1	生別	4	1	5.59	6.32	4.09
	2	内浦	1	3	-0.03	-0.03	-0.14
	3	同属	1	2	0.03	0.03	0.22
	4	ジェニングス	1	1	0.39	0.43	0.02
	5	リゲニン	3	2	2.07	2.34	1.07
	6	コンポスト	5	1	7.66	8.61	6.17
	7	専焼	2	1	1.82	2.07	0.64
	8	端材	2	1	1.82	2.07	0.64
	9	積丹	2	4	0.05	0.06	0.05
	10	北の丸	2	5	0.00	0.00	-0.15
低頻度と高頻度	1	紅茶	2	153	-88.29	-57.74	-56.40
	2	敬意	2	112	-61.47	-41.05	-39.74
	3	先程	2	247	-150.72	-96.05	-94.68
	4	振る	1	642	-426.93	-259.57	-258.17
	5	大将	1	376	-245.95	-151.03	-149.64
	6	懐	1	197	-124.73	-78.00	-76.62
	7	弦	2	139	-79.09	-52.04	-50.71
	8	ソーシャル	1	115	-69.68	-44.56	-43.20
	9	ぼんやり	1	332	-216.08	-133.08	-131.69
	10	思い出す	1	1343	-905.24	-545.63	-544.22

図9 低頻度と低頻度

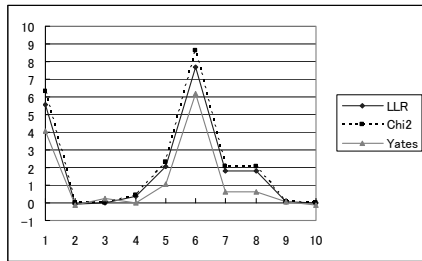
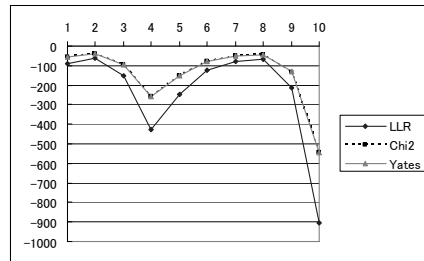


図10 低頻度と高頻度



4. 4 コーパスサイズの違いで現れる特徴

Dunning (1993) は、LLRがテキストサイズが大きくても、小さくても十分に妥当な値を示す指標だと述べている。2×2の分割表に1億語規模のデータを用いることもある特徴語抽出の場合、統計量が非常に高くなるのが問題とされている (高見2009,P4)。そこで、最後の調査では、参照コーパスの総語数を1億語と換算し、対象コーパスの総語数を10万語、100万語、1000万語に換算した場合、3指標の統計量の高さが相対的にどのような関係になるのか、また、その相対的な関係に変化が見られるかを調査した。

この調査では、統計量の高さによって異なる結果が得られると考えたため、4.1で計算した平均統計量のうち、統計量が100以上である高統計量群と100以下の低統計量群に分け、その違いを見ることにした。統計量が100以下という条件を設定した理由は、4.2において統計量が100以下になると、Yatesが低下する現象が見られたためである。次に、各群から、それぞれから各10語無作為抽出を行った。ただし10万語換算での頻度が1以下になるものは値が小さすぎるため、ここでは対象外とした。表6に無作為抽出した20語の頻度及び参照コーパスの総語数を1億語に換算し、対象コーパスの総語数を10万語、100万語、1000万語と換算した場合の統計量を示す。また、表6のIDを横軸にし、統計量を縦軸にしたグラフとして、図11から図16を示す。以下では、視覚的に統計量の相対的な関係を見るため、図を中心に分析する。

まず、10万語換算したデータである図11と図12では、高統計量群と低統計量群では異なった特徴がある。高統計量群では、LLRの統計量が最も低く、Yatesのほうがやや低いながらも、Chi2と類似した高さを示している。それに対し、低統計量群では、Chi2>LLR>Yatesとなっている。高い統計量でLLRが他の2指標よりも低くなる現象、及び低い統計量でYatesが低くなる現象は4.2で得られた結果と同じだといえる。

次に100万語換算したデータを見てみる。図11の傾向は図13でも同じように見られ、その相対的な関係には変化がないが、図12の傾向は、図14とは異なっている。図14では、LLRが最も低く、Yatesのほうがやや低いながらも、Chi2と類似した高さを示している。また、この傾向は図16においても見られる。

これらの結果から、低統計量の場合、Yatesはコーパスサイズによって影響を強く受けていることがわかる。つまり、Yatesはコーパスサイズが小さい10万語レベルでは、統計量が最も低かったが、100万語、1000万語ではChi2と同レベルの統計量を示しており、コーパスサイズに最も影響を受けた指標だといえる。

日本語教育語彙を選定するための統計的指標
 — 尤度比検定、カイ2乗検定、イエーツの補正公式の特徴 —

表6 対象コーパスのサイズ変化と統計量

	I D	見出し語	白書	書籍	10万語換算			100万語換算			1000万語換算		
					LLR	Chi2	Yates	LLR	Chi2	Yates	LLR	Chi2	Yates
高統計量群	1	超える	927	887	12.6	18.0	16.5	123.9	176.2	174.7	1099.9	1438.8	1437.5
	2	指標	444	268	11.8	20.9	18.0	115.7	201.6	198.6	1002.8	1492.0	1489.6
	3	居住	538	276	17.0	32.8	29.1	167.2	314.1	310.5	1434.4	2227.4	2224.6
	4	軽	246	122	8.0	15.8	12.1	79.1	151.0	147.2	677.0	1060.4	1057.5
	5	合わせる	936	1481	3.2	3.7	3.2	31.5	36.7	36.2	285.6	323.0	322.5
	6	土地	2293	1546	53.1	89.4	86.8	522.7	865.8	863.2	4558.0	6580.9	6578.8
	7	不足	543	623	5.0	6.7	5.6	49.9	65.8	64.7	446.2	553.8	552.8
	8	勤労	549	165	27.4	70.0	63.1	267.8	652.7	646.0	2211.3	3888.4	3884.1
	9	賠償	308	257	5.3	8.0	6.2	51.9	78.3	76.4	458.0	623.1	621.4
	10	終了	462	310	10.8	18.2	15.6	105.9	175.9	173.3	923.6	1335.3	1333.1
低統計量群	1	解体	103	214	0.1	0.1	0.0	0.5	0.6	0.4	4.9	5.2	5.0
	2	妨害	66	134	0.0	0.0	0.1	0.4	0.5	0.3	4.1	4.3	4.1
	3	映像	188	265	1.0	1.2	0.6	9.8	11.9	11.2	88.0	103.0	102.3
	4	キロメートル	170	148	2.7	4.1	2.5	26.8	39.6	37.9	236.6	318.0	316.5
	5	差異	102	154	0.4	0.5	0.1	4.1	4.9	4.3	37.5	43.0	42.4
	6	宇宙	378	616	1.1	1.3	0.9	11.2	12.9	12.4	101.3	113.6	113.2
	7	欧米	350	430	2.7	3.5	2.6	27.2	34.9	33.9	243.7	296.5	295.6
	8	潜在	134	193	0.7	0.8	0.2	6.5	7.8	7.1	58.4	67.9	67.2
	9	関心	467	971	0.2	0.3	0.1	2.4	2.6	2.4	22.2	23.2	23.0
	10	生かす	381	541	2.0	2.4	1.7	19.3	23.4	22.7	174.2	203.2	202.6

図 11 10万語レベル (高統計量群)

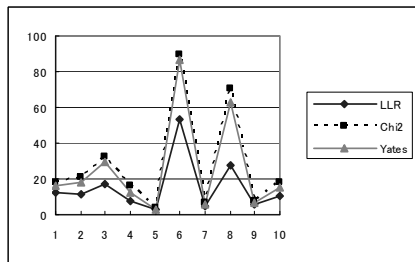


図 12 10万語レベル (低統計量群)

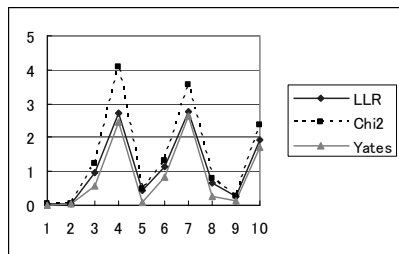


図 13 100万語レベル (高統計量群)

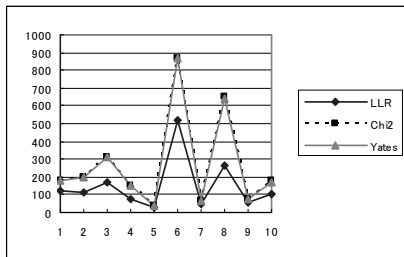
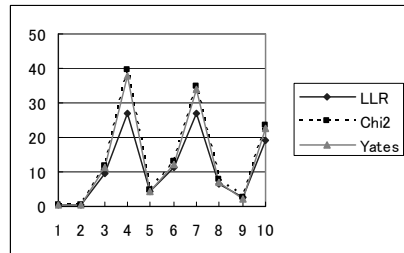


図 14 100万語レベル (低統計量群)



5. 考察

以上、本稿では、1) 各指標に基づき抽出された特徴語の違い、2) 統計量の高さの相対的關係、3) 低頻度語における統計量の高さ、4) コーパスサイズの統計量への影響という4点について調査してきた。以下、調査でわかったことをまとめる。

- 1) 非共通語の数での調査では、LLRとChi2の指標間で共通性が高く、LLRとYatesの指標間で共通性が低かったといえる。しかし、それは抽出の範囲によっても大きく異なる。また、各指標間で1-8000位までを対象に抽出すると、その指標間の非共通語は非常に少なくなったことから、指標及び抽出範囲は語彙選定に影響を与える要因であるといえる。
- 2) 統計量が高いケースでは、Chi2とYatesは他の2指標よりも統計量が高くなったが、LLRは統計量が低くなった。一方、統計量が高いケースでは、Yatesは他の2指標よりも統計量が低くなったが、Chi2は統計量が高いままであった。
- 3) 低頻度同士で統計量を計算したケースでは、Yatesが最も統計量を抑えたが、低頻度と高頻度で統計量を計算したケースでは、LLRが最も統計量を抑えた指標であった。
- 4) 低統計量の場合、Yatesは10万語レベルで統計量が最も低かったが、100万語、1000万語ではChi2と同レベルの統計量を示しており、コーパスサイズに最も影響を受けやすい指標であった。

以上の4点の調査の結果に基づき、それぞれの指標の特徴を考察したい。まず、LLRは他の2指標よりも安定的に統計量を抑える指標であると考えられる。統計量が高いケースでは、他の2指標よりも統計量を抑えることができ、低頻度語と高頻度を比べたケースでは、他の2指標よりも統計量が低かった。Yatesと比べ、コーパスサイズに大きな影響を受けなかったことも安定的な指標だといえる要因である。また、非共通語の調査からは、Yatesとの共通性が低く、Chi2との共通性が高い傾向があることがわかった。

Yatesは、統計量が高いケースでは、LLRよりも統計量が高くなり、Chi2と同レベルの統計量を示すが、統計量が低くなるにつれ、統計量が他の2指標よりも低くなるといえる。低頻度語同士を比べる場合にも最も統計量を抑えられるが、コーパスサイズに影響を受けやすいというマイナス面もあるため、安定性が低く、使い方に注意が必要な指標だと思われる。

Chi2は統計量が高いケースや低いケース、低頻度語でのケース、コーパスサイズを変えた場合のケースといったように、どの調査においても、他の2指標よりも統計量が高くなる傾向があることがわかった。統計量が高いということは、その語を特徴語として高く評価する可能性があるため、高すぎる統計量が問題となる特徴語の抽出において、Chi2が安定的に統計量を抑えられるLLRよりも優れた指標だとは言い難いと考えられる。

6. 今後の課題

本稿では今後、日本語教育語彙の選定を行うために必要となると考えられ、特徴語を抽出することができるLLR、Chi2、Yatesの3つの指標にどのような違いがあるのかを考察してきた。先行研究での報告が少ないなか、本稿で示した各指標の特徴は今後、特徴語の抽出及び指標を選択するケースにおいて、参考になるのではないだろうか。しかし、日本語教育語彙を選定する場合、この指標だけを用いることはないと思われる。LLR、Chi2、Yatesといった特徴度を示す指標は、あくまで語彙選定をするための一つの目安にしかならない。実際の語彙選定では、語が使用されるレンジ、難易度、新密度といった様々な指標も考え、総合的に選定する必要がある。本稿では、そういった指標の一つを考察したにすぎないため、引き続き、語彙選定のための指標に関する研究、及びそれらをどう語彙選定に活用するのかを具体的に検討していきたい。

注

1. WordSmith は Mike Scott 氏によるコーパス分析ツール
Antconc は早稲田大学の Laurence Anthony 氏によるコーパス分析ツール
2. Chi2及びYatesは稲垣他（1992）のp. 149を参考にしたが、この計算式は一般的な統計書に書かれている内容である。
3. ③は Kilgarriff（1996）によって示された計算式であるが、この計算式は、池田他（1989）のp. 95で示されたものと実質的に同じで、得られる統計量は同じである。
4. 中條他（2004）で使用された計算式は <http://www5d.biglobe.ne.jp/~chujo/> で確認できるが、この計算式で求められる統計量に2を乗じると、Kilgarriff（1996）の計算式で求められる統計量と同じになる。
5. 各サンプルはサンプルの長さを1000字に固定した固定長サンプル、それを章や節など意味のまとまりで区切った可変長サンプルがあるが、本稿では可変長サンプルを用いた。
6. 図1のID6が約12%となり、最大の差となっている。

参考文献

- Dunning, T. (1993). *Accurate methods for the statistics of surprise and coincidence*. *Computational Linguistics*, 19 (1), 61–74.
- Kilgarriff, A. (1996). *Which Words Are Particularly Characteristic of a Text? A Survey of Statistical Approaches*. *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*. Sussex, April 1996, 33-40
- 池田他央（編）（1989）『統計ガイドブック』新曜社
- 稲垣宣生, 山根芳和, 吉田光雄（1992）『統計学入門』裳華房
- 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巻3号, 165-197
- 近藤明日子（2009）「中学校教科書の教科特徴語の抽出と考察「現代日本語書き言葉均衡コーパス」の語彙との比較から」『特定領域研究日本語コーパス 平成20年度公開ワークショップ（研究成果報告会）予稿集』（特定領域研究日本語コーパス総括班）, 117-122
- 大学英語教育学会基本語改訂委員会（編）（2003）「大学英語教育学会基本語リスト JACET List of 8000 Basic Words」
- 高見 敏子（2003）「「高級紙語」と「大衆紙語」の corpus-driven な特定法」『（北海道大学）大学院国際広報メディア研究科言語文化部紀要』44, 73-105
- 高見敏子（2009）「言語データと統計的検定に関する疑問—統計学の入門書でわからないこと」『コーパス言語研究における量的データ処理のための統計手法の概観』統計数理研究所共同研究レポート232, 1-10
- 太郎丸博（2006）『人文・社会科学のためのカテゴリカル・データ解析入門』ナカニシヤ出版
- 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紀要』18号, 99-108
- 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」『英語コーパス学会』12, 19-35
- 橋本直幸（2009）「BCCWJを利用した日本語教育語彙リスト作成の試み」『特定領域研究 日本語コーパス 平成20年度公開ワークショップ（研究成果報告会）予稿集』特定領域研究日本語コーパス総括班, 183-190
- 前川喜久雄, 山崎誠（2009）「現代日本語書き言葉均衡コーパス」『国文学解釈と鑑賞』74巻1号, 至文堂, 15-25
- 丸山岳彦（2009）「現代日本語書き言葉均衡コーパスから見えるもの」『日本語教育』140号, 日本語教育学会, 26-36

使用コーパス

- 国立国語研究所（2008）『現代日本語書き言葉均衡コーパスモニター公開データ（2008年度版）』