語彙調査における高頻度語の分離に関する計量的検討

A Quantitative Study on Separation of High Frequency Words Obtained as Vocabulary Survey Results

石井　正彦
ISHII Masahiko

キーワード：語彙調査　基本語彙　高頻度語　層間変動　特化係数　散布度

要　旨

大規模な語彙調査を行うと，得られた語彙は少数の高頻度語と大多数の低頻度語とに分離することが知られている。高頻度語は基本語彙の候補となることから更なる分離が追求され，主に語の「使用範囲」を利用した基本語彙の抽出が試みられてきた。ただし，使用範囲は，その統計量としての性質上，より精度の高い分離の指標としては十分でない。本稿では，語の使用量の層間変動を新たな指標とし，その大きさを「対数化特化係数散布度」によって測定する方法を提案して，高頻度語がどのように分離されるかを検討する。結果として，現代新聞の(内容による層別を施した)語彙調査を行って得た高頻度上位約 100 語について，対数化特化係数散布度の小さい方から概ね骨組み語→叙述語→テーマ語の順に分布（分離）する傾向を見出し，その有効性を確認した。

１．はじめに：高頻度語の分離問題

十分な規模の言語資料について語彙調査を行うと，語の使用頻度別の異なり語数の分布は，図１の［Ａ］のようにＬ字型を描く。これは，語彙が「使われる回数の少ない数多くの語」と「使われる回数の多い数少ない語」とに大きく分かれる傾向を示している。そのことは，グラフに［Ｂ］のように逆Ｌ字型を描く延べ語数の分布を重ねてみるとわかりやすい。

このように，大規模な語彙調査で得られた語彙は少数の高頻度語と大多数の低頻度語とに分離することが知られている。さらに，高頻度語には基本語彙の

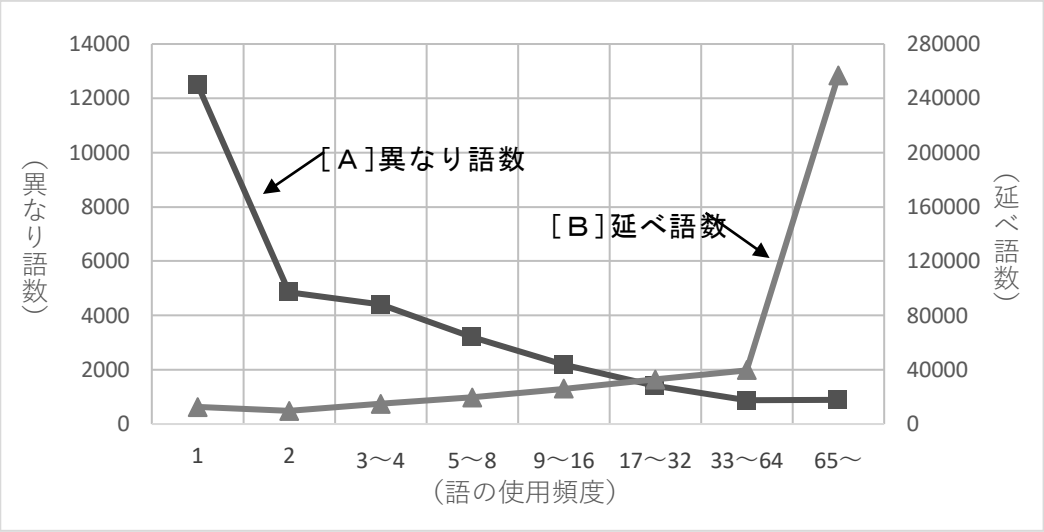

図１　語の使用頻度別の度数分布

（国立国語研究所(1964:58-59)「表 2.5」「表 2.6」から作成）

候補となる語が含まれており，計量語彙論では，高頻度語から基本語彙を取り出す，言い換えれば，高頻度語を基本語彙とそうでないものとに分離する方法が検討・開発されてきた。それは，基本的には，層別を行った単独の語彙調査において，あるいは，複数の異なる語彙調査の比較において，「使用率」の大きい語のうち，出現した層や調査の数すなわち「使用範囲」が大きい（広い）ものを基本語彙の要素とするというものである。たとえば，林四郎(1971)は，国立国語研究所の「新聞の語彙調査」（1966 年）のデータを用い，「話題」による層別 12 区分（政治，外交，経済，社会，文化等の紙面別に相当）を用いて，そのうちいくつの層に出現したかを「広さ」と表し，広さ１～３を「狭い」，４～６を「中位」，７～９を「かなり広い」，10～12 を「極めて広い」と４区分し，後二者に属する語を（使用率を反映した「深さ」との関係も考慮して）基本語彙（林の用語では「基幹語彙」）としている。

ただし，この例からもわかるように，語の「使用範囲」は，語彙調査で設定された層の数（上の例では 12）を上限とし，また，語の使用頻度と無関係に，ある層に出現すれば１，出現しなければ０とカウントされて合計されるため，個々の語の値の範囲が限られ（上の例では１～12），また，同じ値を持つ語が多数生じるなど，（高頻度語から基本語彙を分離・抽出するには有効であっても）

より精度の高い分離の指標としては十分でない。そもそも，高頻度語についてそれがなぜ高頻度であるのか，その理由を問うことは，基本語彙の抽出・選定とは別の，計量語彙論の最も重要な課題の一つである。そして，その問いに答えていくためには，高頻度語をより高い精度で分離し，得られた結果について質的な検討を進めていくことが必要である。そこで，本稿では，高頻度語の分離問題を計量語彙論の基本的な研究課題とする立場から，高頻度語を計量的に有効な精度で分離し得る新たな指標を提案し，現代新聞の語彙調査から得た高頻度語を例にその有効性を試行的に検討してみたい。

２．語の使用量の層間変動の大きさ

一般に，計量語彙論における「基本語彙」は，次のように定義される。

> 基本語彙とは，使用率が大きく，しかも対象とする言語作品あるいは言語体系の中に幾つかの層を設けて考えることができる場合（たとえば雑誌であれば，実用記事・文芸作品・趣味など掲載する内容別の層を設け，また平安時代物語であれば作品別に層を設けることができる），できるだけ多くの層にわたって出現する語の集合をいう。（樺島忠夫 1980：345）

ここで「使用率」と「使用範囲」（出現する層の数）とは一応独立した統計量と想定されるが，実際には両者の間には高い相関が見られ，高頻度語の使用範囲は概して大きく（広く）なる。高頻度語を分離しようとするとき，使用範囲が有効な指標となりにくい理由はこのようなところにもある。

では，高頻度であり，かつまた，使用範囲も大きいという語群において，それでもなお語によって異なり得る統計量とは何か。石井（2015）は，語がすべての層を通してどれほど均等に用いられているかを測る指標として「語の使用量の層間変動の大きさ」に注目しているが，これもそうした指標の一つと言えるだろう。つまり，すべての層に出現するような高頻度語でも，各層における使用頻度（使用率）のばらつきの幅には差があり得ると考えられるのである。たとえば，石井（2015）で用いた，国立国語研究所「高校教科書の語彙調査」（1974 年の理科・社会科 9 科目を層別とする全数語彙調査）の「W単位」語彙表（国立国語研究所 1984）で，全体順位４位の「この」と６位の「その」は，いずれも全９科目に出現している（したがって使用範囲はともに９）高頻度語であるが，

それぞれの各科目での使用率をみると(表１),「この」の方が「その」よりも科目間のばらつきの幅が大きい。そのことは，これをグラフにしてみるとさらにわかりやすい(図２)。

表１　「この」「その」の科目別使用頻度と使用率（カッコ内，‰）

語	物理	化学	生物	地学	倫理社会	政治経済	日本史	世界史	地理Ｂ
(4)「この」	595 (25.1)	257 (13.1)	290 (13.0)	186 (10.9)	315 (11.0)	359 (11.2)	338 (9.2)	273 (9.1)	105 (4.5)
(6)「その」	302 (12.7)	195 (10.0)	207 (9.3)	165 (9.7)	259 (9.0)	483 (15.1)	457 (12.4)	189 (6.3)	184 (7.8)
延べ語数	23707	19551	22374	17044	28742	32086	36828	30051	23472

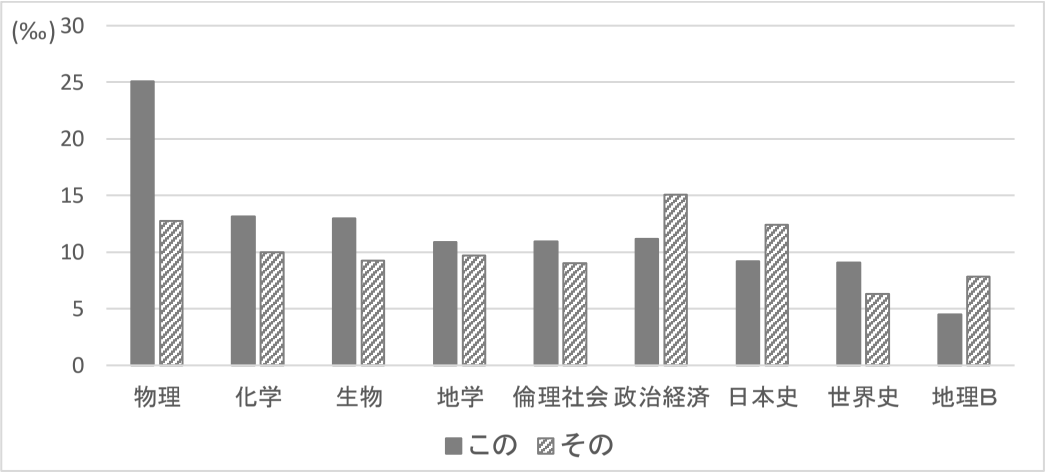

図２　「この」「その」の科目別使用率（‰）

「この」の使用量は「物理」でとくに多く[1)]，また,「地理Ｂ」でとくに少ないのに対し,「その」にはこうした顕著な使用・不使用の傾向を示す科目は見当たらない。同じ高頻度語，しかも，同じ指示連体詞というよく似た語であっても,その使用量の層間変動の大きさには少なからぬ違いが見出せるのである(両語の分布の違いは，これを標本データとみなしてカイ二乗検定をすると１％水準で有意（$\chi^2(8)=183.390$，$p<.01$)といえる差である)。

本稿では，以降，この「語の使用量の層間変動の大きさ」について，高頻度語分離の指標として有効かどうかを検討する。その統計量は，上の例のように

「使用率」によっても測定できるが，語の層別使用率はその層の延べ語数・異なり語数の違いの影響を受け，また，変動の幅を評価することも簡単ではないため，石井(2015)では「対数化特化係数散布度」という統計量を提案している。

「特化係数」とは，表２・３に示すように，全体（語彙量Ｖ）の構成比（内訳）を基準にして，各項目（語W）の構成比を，その何倍にあたるかという形で表したもの（構成比の相対比）である（上田 1981）。通常，語は層の大きさ（言語量）に応じてその使用頻度を変えるが，もし語Wが各層（S_1～S_4）の大きさ以外の違いに影響されなければ，その構成比は語彙量Ｖの構成比と一致し，両者の比すなわち特化係数はいずれの層においても１になる。しかし，もし語Wが層の大きさ以外の違いに影響され，たとえば層 S_1 ではその使用が促進され，層 S_2 では逆に阻害されたとすれば，S_1 における語Wの特化係数“$(f_1/F)/(t_1/T)$”は１より大きくなり，S_2 における特化係数“$(f_2/F)/(t_2/T)$”は１より小さくなる（最小値は０）。

表２　語彙表におけるWとＶの頻度

語＼層	S_1	S_2	S_3	S_4	計
：					：
：					：
W	f_1	f_2	f_3	f_4	F
：					：
：					：
V	t_1	t_2	t_3	t_4	T

表３　Wの特化係数

語＼層	S_1	S_2	S_3	S_4
：				
：				
W	$\frac{f_1/F}{t_1/T}$	$\frac{f_2/F}{t_2/T}$	$\frac{f_3/F}{t_3/T}$	$\frac{f_4/F}{t_4/T}$
：				
：				

ただし，特化係数は，基準値より小さい場合は０～１の間に，大きい場合は１～＋∞に分布して対称的でないため，散布度を求めるには対数変換をする必要がある。その際，値が０の場合には変換できないため，各語の使用頻度に始数（適当な小さい正の数）を加えて特化係数を再計算する。対数化した特化係数のばらつき（散布度）は，基準値０を平均値とみなした標準偏差か，最大値と最小値の差であるレンジ（最大値－最小値）で表し得るが，石井(2015)ではレンジを採用している。

ここで，上の「この」と「その」を例として，それぞれの対数化特化係数散布度を求める手順を示すと表５・６のようになる。表４は，特化係数を求める際に基準値とする９科目（層）の構成比を，各科目のすべての語の使用頻度に

表４　基準となる各科目の延べ語数と構成比

	物理	化学	生物	地学	倫理社会	政治経済	日本史	世界史	地理Ｂ	全体
(A) 各語の頻度の計	23707	19551	22374	17044	28742	32086	36828	30051	23472	233855
(B) 各語の「頻度+始数」の計	28,234.9	24,078.9	26,901.9	21,571.9	33,269.9	36,613.9	41,355.9	34,578.9	27,999.9	274606
(C) 構成比	0.103	0.088	0.098	0.079	0.121	0.133	0.151	0.126	0.102	1.000

表５　対数化特化係数散布度を求める過程（「この」の場合）

	物理	化学	生物	地学	倫理社会	政治経済	日本史	世界史	地理Ｂ	全体
①頻度	595	257	290	186	315	359	338	273	105	2718
②頻度+始数(1/9)	595.111	257.111	290.111	186.111	315.111	359.111	338.111	273.111	105.111	2719
③構成比	0.219	0.095	0.107	0.068	0.116	0.132	0.124	0.100	0.039	1.000
④特化係数（②÷(C)）	2.129	1.078	1.089	0.871	0.957	0.991	0.826	0.798	0.379	
⑤特化係数の対数	0.328	0.033	0.037	-0.060	-0.019	-0.004	-0.083	-0.098	-0.421	
⑥最大値	0.328									
⑦最小値									-0.421	
⑧レンジ（⑥－⑦）	0.749									

表６　対数化特化係数散布度を求める過程（「その」の場合）

	物理	化学	生物	地学	倫理社会	政治経済	日本史	世界史	地理Ｂ	全体
①頻度	302	195	207	165	259	483	457	189	184	2441
②頻度+始数(1/9)	302.111	195.111	207.111	165.111	259.111	483.111	457.111	189.111	184.111	2442
③構成比	0.124	0.080	0.085	0.068	0.106	0.198	0.187	0.077	0.075	1.000
④特化係数（②÷(C)）	1.203	0.911	0.866	0.861	0.876	1.484	1.243	0.615	0.739	
⑤特化係数の対数	0.080	-0.040	-0.063	-0.065	-0.058	0.171	0.094	-0.211	-0.131	
⑥最大値						0.171				
⑦最小値								-0.211		
⑧レンジ（⑥－⑦）	0.382									

始数 1/9 を加えたものの総計を延べ語数として求めたものである。

表５・６では，「この」「その」それぞれについて，同じようにして９科目の構成比を求め（③），それを表４の各科目の構成比で除して各科目の特化係数を算出（④），さらに対数変換して（⑤），得られた９科目の最大値と最小値の差（レンジ）を対数化特化係数散布度として求めている。結果は，「この」が 0.749，「その」が 0.382 となって，前者の方が「使用量の層間変動の大きさ」が大きいという，先に使用率の層間変動を見たときと同じ結果となった。次節では，この「対数化特化係数散布度」が高頻度語分離の指標としてどの程度機能するか，新聞における高頻度語の分離を例に検討する。

３．対数化特化係数散布度による新聞高頻度語の分離

「内容にもとづく層別を施した語彙調査」を行い，得られた高頻度語について「その使用量の層間変動の大きさ」を「対数化特化係数散布度」（以下，単に「散布度」とすることがある）により測定して一次元に並べたとき，そこに何らかの分離の傾向が見出せるかを検討する。

語彙調査には，2010 年の『毎日新聞』から抽出した標本データを用いる。具体的には，『ＣＤ－毎日新聞 2010 データ集』（毎日新聞社の許諾を得て購入・使用）から毎月 3 日分（5 日・15 日・25 日），計 36 日分の朝刊（全国版）全紙面の記事（見出し・本文）を抽出し，言語量の大きい８紙面（一面・二面・社説・経済面・国際面・社会面・家庭面・スポーツ面）を層別として，その自立語データを用いる。ここでいう自立語データとは，対象とする紙面のテキストについて，日本語形態素解析システム“unidic-mecab”[2)]により「短単位」に解析し，それを自作のプログラムにより（自立語相当の）「擬似長単位」に再解析して，助詞・助動詞・記号の類を除いたものである。各紙面（層）の語彙量は，表７のようになる。

表７　2010 年『毎日新聞』８紙面の語彙量（擬似長単位）

	一面	二面	社説	経済	国際	社会	家庭	スポーツ	計
延べ語数	30653	36576	18295	62321	40777	73520	75856	144303	482301
異なり語数	12454	12871	6919	21886	14887	26057	21649	34045	102051

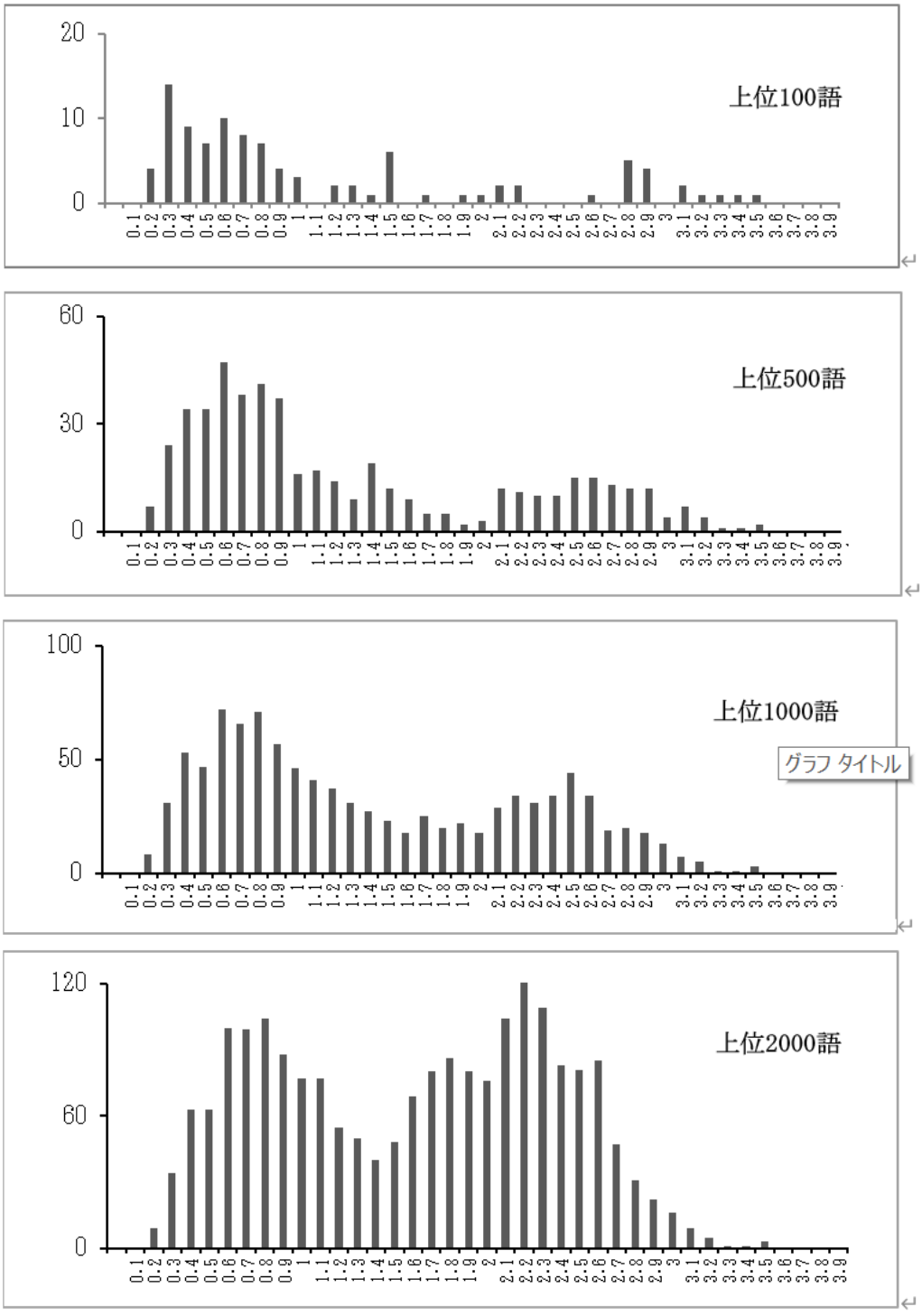

図３　新聞高頻度語の散布度の分布

（横軸：対数化特化係数散布度，縦軸：語数）

いま，表 7 の異なり約 10 万語についてその対数化特化係数散布度を求め，使用頻度順上位 100 語，500 語，1000 語，2000 語のそれぞれについて，散布度 0.1 を階級幅としてその度数分布をグラフに描くと前ページの図 3 のようになる。縦軸のスケールが異なることに注意しつつこれらを比べると，上位 100 語から 2000 語へと高頻度語の範囲を広げていくにつれて，グラフはほぼ双峰性の分布形となり，新聞の高頻度語が散布度の小さい語群と大きい語群とに分離する様子が見てとれる。ただし，より丁寧に見ると，たとえば上位 100 語から 1000 語までのグラフでは左側の山にさらに 2 つの峰があるように見えるし，上位 2000 語のグラフでは右側の山にさらに 2 つないし 3 つの峰があるようにも見える。したがって，対数化特化係数散布度によって高頻度語がいくつかの語群に分離する確度は高いとしても，そこにどのような有意味な類別が見出せるかについては，具体的な語の分布の様子を調べる必要がある。そこで，以下では，語数は少なくなるが，最も高頻度のグループである上位 100 語の分布について検討することにしたい。

次ページから始まる表 8 に，頻度順上位 100 語のうち，本来の新聞紙面にない著作権関係の文言を構成する 3 語を除いた 97 語について，対数化特化係数散布度の昇順に配列した上で，各語の各層（紙面）における（対数化する前の）特化係数を示した。なお，表中の品詞分類は数詞・記号以外は“unidic-mecab”に従うが，見出し語の表記には“unidic-mecab”によるものをわかりやすく変えたものがある。

すでに述べたように，特化係数とは，ある語のある層における構成比（その語の総使用頻度に対するその層での使用頻度の割合）がその層全体の構成比（語彙調査で得られた延べ語数全体に対するその層の延べ語数の割合）を基準値としてその何倍であるかを示す統計量（構成比の相対比）である。その「評価」に絶対的な基準はないが，一般に，特化係数が 2.0 以上，すなわち，基準値（全体の傾向）の 2 倍以上使われていれば「著しい（偏った）使用」，特化係数が 0.5 以下，すなわち，基準値の 1/2 以下しか使われていなければ「著しい（偏った）不使用」であると判断することが多い（上田 1981：33）。逆に言えば，特化係数が 0.5 から 2.0 の間にあるなら，その層でのその語の使用量はほぼ「期待どおり（普通）」と判断されることになる。

表８　新聞高頻度上位 97 語の層別特化係数

群	見出し語	品詞	一面	二面	社説	経済	国際	社会	家庭	スポーツ	散布度
	強い	形容	0.964	1.436	1.214	1.056	1.263	0.969	0.858	0.816	0.245
	なる(成)	動詞	0.833	1.061	1.170	1.147	0.815	0.887	1.478	0.780	0.277
	この	連体	0.948	0.895	1.103	0.672	0.823	1.052	1.281	1.055	0.280
	なか	名詞	0.721	1.045	1.085	1.307	1.160	0.712	1.363	0.800	0.282
	来る	動詞	1.057	0.730	1.148	0.897	0.942	0.951	1.510	0.849	0.316
	これ	代名	1.221	1.260	1.424	0.730	1.491	0.682	1.299	0.749	0.340
	うえ	名詞	1.320	1.373	0.880	0.987	0.998	0.999	1.368	0.622	0.344
	高い	形容	1.177	0.885	1.015	1.220	0.881	1.018	1.454	0.656	0.345
	出す	動詞	0.883	0.913	0.968	0.617	0.711	1.404	1.309	0.949	0.357
	取る	動詞	0.826	1.027	1.509	0.728	1.136	0.658	1.357	1.011	0.360
	する(為)	動詞	1.014	1.047	1.250	1.060	1.059	1.303	1.280	0.564	0.364
	ひらく	動詞	1.026	1.348	0.684	1.314	1.239	1.378	0.864	0.587	0.370
	受ける	動詞	1.096	1.381	0.629	1.041	1.239	1.445	0.962	0.607	0.377
A	また	接続	1.179	1.033	1.083	1.243	1.367	1.130	1.084	0.570	0.380
	ない	形容	1.043	1.082	1.625	0.674	0.878	1.017	1.554	0.719	0.382
	持つ	動詞	1.073	0.878	0.814	0.774	0.721	0.990	1.787	0.818	0.394
	続く	動詞	1.073	0.940	0.596	1.355	1.327	0.544	0.743	1.191	0.397
	出る	動詞	0.931	1.232	0.524	0.715	0.889	1.048	1.312	1.009	0.399
	ため	名詞	0.577	0.737	0.558	1.522	1.038	0.758	1.312	0.976	0.436
	日本	名詞	0.754	1.093	1.243	1.344	0.886	0.489	0.684	1.322	0.439
	いう	動詞	1.128	0.866	1.425	0.593	0.828	1.338	1.642	0.632	0.442
	もの	名詞	1.049	1.078	1.144	0.805	0.959	0.808	1.853	0.669	0.443
	こと	名詞	1.092	1.134	1.310	1.039	0.802	1.009	1.619	0.571	0.452
	いる(居)	動詞	0.901	0.978	0.952	0.897	1.373	1.360	1.473	0.506	0.464
	入る	動詞	0.722	0.719	0.468	0.635	0.740	1.101	1.150	1.384	0.471
	ある	動詞	1.146	0.986	1.669	0.806	1.014	1.080	1.567	0.563	0.472
	その	連体	1.326	0.788	1.405	0.559	0.799	0.730	1.687	0.939	0.480
B	しかし	接続	0.671	1.430	2.162	1.052	0.971	0.636	1.429	0.678	0.531
A	それ	代名	1.213	0.855	1.718	0.796	0.638	0.548	1.863	0.831	0.531
	後(あと)	名詞	0.902	0.427	0.691	0.514	0.922	0.977	1.034	1.511	0.549
	決める	動詞	0.881	0.613	0.519	1.123	0.466	0.900	0.497	1.699	0.562
	よう(様)	形状	0.926	0.777	1.512	0.723	0.754	1.070	2.078	0.562	0.568
	行く	動詞	0.813	1.047	1.012	0.841	0.482	0.878	1.867	0.863	0.588
	言葉	名詞	1.393	1.325	0.400	1.586	0.835	1.102	1.321	0.451	0.598
	語る	動詞	1.193	2.002	0.895	0.869	0.828	1.392	0.495	0.848	0.607
	目指す	動詞	0.583	1.368	0.878	1.736	1.141	0.647	0.420	1.144	0.616
	今回	名詞	1.163	1.309	1.941	1.229	1.770	1.046	0.680	0.470	0.616
	行う	動詞	0.710	0.524	0.557	0.532	1.157	0.849	0.431	1.879	0.639
	可能性	名詞	1.242	1.329	0.882	1.472	1.721	1.652	0.383	0.381	0.654
B	対する	動詞	1.404	1.887	1.168	1.310	1.793	0.982	0.606	0.405	0.668
	声	名詞	1.180	1.721	0.662	1.160	1.181	1.360	1.186	0.357	0.683
	多い	形容	0.781	0.962	0.780	1.069	0.884	0.828	2.317	0.475	0.689
	おる(居)	動詞	1.000	1.282	0.762	1.860	1.864	1.193	0.599	0.381	0.689
	上げる	動詞	0.723	0.878	0.685	0.594	0.389	0.595	0.790	1.907	0.690
	今	名詞	0.922	0.780	1.075	0.507	0.902	0.830	2.625	0.523	0.714
	見る	動詞	0.997	0.945	0.736	0.879	1.341	1.714	1.537	0.320	0.728
	出来る	動詞	0.752	0.739	0.840	0.921	0.405	0.911	2.167	0.809	0.729
	つく	動詞	1.393	1.948	1.337	1.507	1.270	1.148	0.596	0.353	0.742
	分かる	動詞	0.851	0.686	0.629	0.581	0.576	2.348	1.698	0.421	0.746

群	見出し語	品詞	一面	二面	社説	経済	国際	社会	家庭	スポーツ	散布度
	日(ひ)	名詞	0.660	0.308	0.443	0.288	0.314	1.204	1.451	1.614	0.748
	問題	名詞	1.099	1.819	2.060	1.048	1.124	1.291	0.928	0.336	0.787
	国(くに)	名詞	1.778	2.039	1.971	0.940	0.991	0.800	1.167	0.332	0.789
	求める	動詞	1.258	1.962	1.233	1.251	1.565	1.390	0.573	0.309	0.803
B	使う	動詞	0.633	0.620	0.635	1.373	0.719	1.247	2.282	0.350	0.814
	どう	副詞	0.897	1.475	2.284	0.899	0.665	0.745	2.084	0.329	0.842
	因る	動詞	1.070	0.886	0.558	1.157	1.892	2.054	0.757	0.289	0.852
	考える	動詞	0.916	0.772	1.199	0.682	0.371	0.853	2.655	0.569	0.855
	良い	形容	0.519	0.623	0.518	0.329	0.279	0.787	2.103	1.395	0.877
	何	代名	0.943	0.587	1.724	0.315	0.508	0.912	2.394	0.761	0.881
C	示す	動詞	1.274	2.609	1.499	1.807	1.490	0.524	0.325	0.428	0.905
B	必要	名詞	1.299	1.946	2.412	1.069	0.791	0.586	1.638	0.265	0.959
	めぐる	動詞	1.556	2.096	1.472	1.001	1.607	1.538	0.462	0.227	0.966
	思う	動詞	0.691	0.478	0.594	0.290	0.323	0.898	2.759	0.954	0.978
	韓国	名詞	0.571	0.968	0.573	0.896	2.369	0.217	0.227	1.658	1.039
	時(とき)	名詞	0.792	0.592	0.709	0.518	0.285	0.783	3.126	0.632	1.040
	話す	動詞	0.651	0.434	0.162	0.486	0.507	1.928	1.822	0.870	1.075
	自分	名詞	0.362	0.255	0.158	0.247	0.205	0.871	2.866	1.209	1.259
	発表する	動詞	0.539	0.874	0.221	2.707	1.236	1.125	0.137	0.840	1.297
C	米国	名詞	0.876	1.049	1.046	1.626	3.316	0.149	0.289	0.800	1.347
	人(ひと)	名詞	1.138	0.641	0.793	0.795	0.606	0.998	3.148	0.139	1.354
	東京	名詞	0.901	0.448	0.061	1.114	0.223	1.720	0.847	1.274	1.447
	私(わたくし)	代名	0.509	0.640	0.139	0.416	0.359	0.486	4.536	0.197	1.513
	中国	名詞	0.805	1.169	1.030	1.664	3.189	0.234	0.095	0.872	1.524
	述べる	動詞	1.877	3.002	0.414	1.214	2.006	1.214	0.089	0.215	1.530
	写真	名詞	0.208	0.063	0.054	1.884	1.676	1.787	1.379	0.396	1.546
	政府	名詞	1.593	3.304	1.621	1.893	1.507	0.395	0.401	0.085	1.590
	1	数詞	0.283	0.227	0.081	0.285	0.118	0.180	0.184	3.145	1.590
	10	数詞	0.140	0.153	0.054	0.236	0.058	0.255	0.183	3.214	1.777
	4日	数詞	0.553	1.436	0.125	1.003	0.668	1.297	0.018	1.651	1.974
	2	数詞	0.263	0.214	0.029	0.229	0.059	0.166	0.244	3.184	2.041
	6	数詞	0.117	0.146	0.026	0.141	0.054	0.107	0.159	3.368	2.115
	民主党	名詞	2.174	4.741	2.969	0.745	0.131	0.373	0.781	0.033	2.156
	共同	名詞	0.063	0.108	0.047	1.099	0.754	0.195	0.016	2.734	2.230
	14日	数詞	0.470	1.392	0.074	1.163	0.863	1.191	0.009	1.618	2.258
	首相	名詞	2.482	4.654	2.726	0.559	1.269	0.365	0.288	0.010	2.648
	サッカー	名詞	0.304	0.032	0.005	0.078	0.191	0.167	0.082	3.350	2.805
D	9	数詞	0.064	0.189	0.005	0.175	0.125	0.152	0.118	3.332	2.844
	8	数詞	0.080	0.093	0.004	0.265	0.044	0.154	0.138	3.329	2.915
	J	記号	0.191	0.200	0.006	0.343	0.003	0.260	0.361	3.060	2.962
	W	記号	0.403	0.003	0.005	0.212	0.350	0.397	0.623	2.788	2.969
	7	数詞	0.212	0.112	0.003	0.372	0.053	0.105	0.203	3.224	2.989
	5	数詞	0.225	0.145	0.002	0.196	0.073	0.126	0.137	3.313	3.150
	小沢氏	名詞	3.259	3.356	3.696	0.218	0.002	1.680	0.054	0.015	3.182
	4	数詞	0.232	0.163	0.002	0.336	0.010	0.103	0.184	3.247	3.252
	24日	数詞	0.593	1.131	0.024	1.125	0.612	1.246	0.001	1.751	3.334
	3	数詞	0.090	0.131	0.001	0.281	0.049	0.177	0.176	3.272	3.446
	0	数詞	0.175	0.031	0.001	0.057	0.005	0.015	0.067	3.550	3.589

これを踏まえ，表８では，特化係数が 2.0 以上のセルと 0.5 以下のセルに網掛けを施し，前者についてはさらに太線で囲んで表示した。その上で改めて表８を眺めると，これら網掛けのセルは，表の上部では（「日本」が社会面でわずかに 0.5 を下回るのを除けば）まったく見られないが，下に行くにしたがって徐々に増え，表の下部ではほとんどが網掛けのセルになるというように，上から下へと段階的に増えていることがわかる。つまり，対数化特化係数散布度（＝語の使用量の層間変動の大きさ）が大きくなれば，その語の「著しい（偏った）使用／不使用」も増える，という傾向が明確に認められるわけである。

そこで，表８の各語における「著しい（偏った）使用／不使用」を示す「偏った層」に注目すると，これら 97 語は，そうした「偏った層」が８層のうちいくつあるかという観点から，おおよそ以下の４群に区分することができる。

Ａ群（偏った層の数が０）

どの紙面でも普通に（基準値に近く）使われる，すなわち８紙面でほぼ均等に使われる語。対数化特化係数散布度は概ね 0.5 以下。

Ｂ群（偏った層の数が１～２）

ほとんどの紙面で普通に使われるが，ある一つの紙面で顕著に使われたり，逆に顕著に使われなかったりするか，あるいはその両方が並存するかする語。散布度は概ね 0.5 以上 0.9 以下。

Ｃ群（偏った層の数が３～５）

ある一つの紙面では顕著に使われるが，別のいくつかの紙面では顕著に使われないという語。普通に使われる紙面もいくつかある。散布度は概ね 0.9 以上 1.6 以下。

Ｄ群（偏った層の数が６～８）

ある一つの紙面で顕著に使われるが，他のほとんどの紙面で顕著に使われない（普通に使われる紙面が少ない）語。散布度は概ね 1.6 以上。

そして，これら４群に所属する語を品詞別に整理すると，次のようになる（品詞名に続くカッコ内の数字は語数，同一品詞内は散布度昇順に配列した）。なお，各群の境界付近では語によって所属が交差する場合も生じたが，その詳細については表８最左端の「群」列を参照されたい。

A 群(28 語)

名　詞(6)：なか，うえ，ため，日本，もの，こと
代名詞(2)：これ，それ
動　詞(14)：なる(成)，来る，出す，取る，する(為)，ひらく，受ける，持つ，続く，出る，いう，いる(居)，入る，ある
形容詞(3)：強い，高い，ない
連体詞(2)：この，その
接続詞(1)：また

B 群(32 語)

名　詞(10)：後(あと)，言葉，今回，可能性，声，今，日(ひ)，問題，国(くに)，必要
代名詞(1)：何
動　詞(16)：決める，行く，語る，目指す，行う，対する，おる(居)，上げる，見る，出来る，つく，分かる，求める，使う，因る，考える
形容詞(2)：多い，良い
形状詞(1)：よう(様)
副　詞(1)：どう
接続詞(1)：しかし

C 群(16 語)

名　詞(9)：韓国，時(とき)，自分，米国，人(ひと)，東京，中国，写真，政府
代名詞(1)：私(わたくし)
動　詞(6)：示す，めぐる，思う，話す，発表する，述べる

D 群(21 語)

名　詞(5)：民主党，共同，首相，サッカー，小沢氏
数　詞(14)：１，１０，４日，２，６，１４日，９，８，７，５，４，２４日，３，０
記　号(2)：Ｊ，Ｗ

なお，この４区分を，図３でみた「上位 100 語」の度数分布のグラフにあてはめると，図４のようになる。

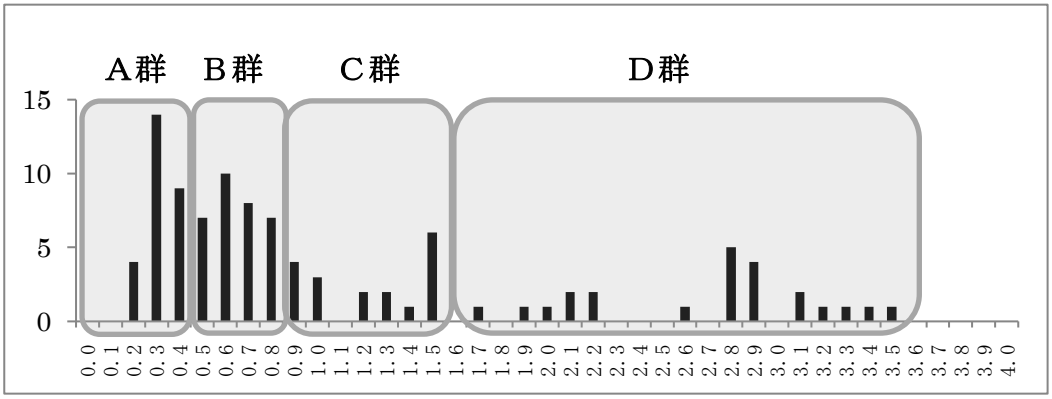

図４　新聞高頻度上位 100 語の散布度の分布と４区分（Ａ～Ｄ群）
（横軸：対数化特化係数散布度，縦軸：語数）

４．分離の傾向：骨組み語・叙述語・テーマ語

では，このＡ～Ｄ群の語に，何らかの違い（分離の傾向）が見出せるだろうか。結論を先取りすれば，ここには，寿岳章子(1967)が提案した高頻度語の３分類(に所属する語)がそれぞれまとまって分布する様子を見ることができる。

寿岳(1967)は，日本語の基本語彙（寿岳の用語では「基礎語彙」）を現代語だけでなく歴史的にも解明していくためには各時代の基本語彙を求める作業が必要になると説き，その一環として「源氏物語の高頻度語」[3)]を「骨組み語・テーマ語・叙述語」の３種に分類している。この３分類は「一つの言語資料内で使われるそれぞれの語がそなえている機能」によるもので,「骨組み語」は,「書くにせよ話すにせよ，とにかく日本語というワク内で言語行動をとるかぎり，絶対に必要なことば」であり,「いつの時代でも，またどんなに質の違う資料でも，どうしても使用してしまうことばのグループ」で,「多くの場合，極めて使用率大という形であらわれる」もの,「テーマ語」は,「その言語資料の内容によって動いてゆく語」であり,「その言語資料が何についての物語であるか，何を主たる素材としているか，あるいは作者のどのようなものの見方によって作られた資料であるか」ということによってその出現・使用が規定されるもの，

「叙述語」は,「その言語資料の叙述方法，あるいは一般に位相などといわれるものが関連して生じることば」で,「たとえばある時代の資料はおしなべてその語をよく使う，またはこの資料群ではよく使われる，あるいはこの作者は何を書いてもこれこれの語を目立ってよく使う,というようなたぐいの語」であり，平安時代の女流文学では副詞の「いと」や敬語動詞のグループなどがその例であるという。

さて，前節のＡ～Ｄ群の語のうち，□で囲んだ語は，林(1971)が国立国語研究所の４種の語彙調査（明治初期文献・婦人雑誌・総合雑誌・新聞３紙）の結果を比較し，そのすべてで 100 位以内にあるとした「基幹度最高第１群」の 12 語と一致するもの，また，□で囲んだ語は,「明治初期文献」調査で文語形が見出し語とされたために「基幹度最高第１群」から漏れたものの，実質的には□と同じ資格を持つと考えてよいもので，いずれも時代や資料の違いを超えて多用される骨組み語の可能性が高いものである。実際，これらはその語彙的意味が抽象的・形式的で，助詞・助動詞などの機能語に近い働きをすると考えられ，骨組み語と呼ぶのにふさわしい語である。そして，これら骨組み語と考えられる語は全部で 15 語あるが，うち 13 語がＡ群に所属しており，Ａ群の半数近くを占めてもいる。

一方，Ａ群と対極にあるＤ群には,「民主党，共同（通信社)，小沢氏」という固有名詞,「首相，サッカー」という具体名詞のほか，数詞と記号が多く所属していて,骨組み語の可能性がある語は一つもない。固有名詞は,寿岳(1967:27)も言うように，典型的なテーマ語であり，具体名詞の「サッカー」はスポーツ面の,「首相」は内政を話題とする二面を中心に一面や社説の，それぞれテーマ語であると考えられる。数詞と記号も，表８を見ると「４日，１４日，２４日」を除いてすべてスポーツ面で多用されており，得点やチーム名などを表すのに使われるテーマ語と考えてよい。このように，Ｄ類はそのほとんどがテーマ語であると考えられる。

これに対して，Ｂ類・Ｃ類には,「何，よう(様)，どう，しかし」「人(ひと)，私(わたくし)」のような骨組み語と考えられる語や,「韓国，米国，東京，中国，写真，政府」のようなテーマ語としてよい語も所属しているが，多くは，抽象的な語彙的意味のもとで新聞の叙述を組み立てることに働く，すなわち，新聞

の叙述語と思われる語が分布している。その働きにはさまざまなものがあるため，単純に整理することは難しいが，たとえば，Ｂ群の「言葉，声」という名詞は，特定の〈言語〉や〈音声〉といった具体的な意味よりも，〈発言内容〉という抽象的な意味で使われ，その〈発言内容〉を具体的に表す節や句を受けたり指示したりする働きによって新聞の叙述を組み立てることが多いものと考えられる。一方で，同じＢ群の「語る」やＣ群の「話す，述べる」はそうした〈内容〉を〈発言する〉ことを表す動詞であり，Ｃ群の「示す，発表する」，Ｂ群の「考える」とＣ群の「思う」なども，まとまった〈内容〉を〈表示・表明する〉〈思考する〉ことを表す動詞である。これらはいずれも，〈内容〉を表す節や節構造，句などを受ける述語として（引用的にも）働くことにより，新聞の叙述を組み立てることが多いと考えられる。

このほかにも，Ｂ群の「可能性，問題」は本来の抽象名詞であるが，これらも節や句で表現された〈内容〉を受けたり指示したりすることが多い。この「問題」という語については，高崎みどり(2012)が McCarthy(1991=1995)の「談話構成語」(discourse-organizing words）の概念を用いて分析している。談話構成語とは，言語の閉じた体系に属して文法的な意味を担う機能語（文法語・虚語とも）と，開いた体系に属して言語の創造的な側面を担う内容語（語彙語・実語とも）との中間にあって，双方の性質を備えているような語であり，テクストの「分節」(テクスト内で先行または後続する意味的なまとまり）を代名詞のように指示したり，より大きなテクストパターンを示してその全体像を予測させたりといった働き，つまり，テクストの内容や分野を伝えるのではなく，テクストに構成と構造を与える働きを担う，というものである（McCarthy 1991=1995:105-107)。私見によれば，この「談話構成語」と「叙述語」とはきわめて近い概念内容をもつと考えられる。

なお，叙述語と思われる語はＡ群にも見ることができる。骨組み語以外の動詞には，「(勧告を）出す，(措置を）取る，(会議を）開く，(指摘を）受ける，(疑いを）持つ，(反発が）続く，(影響が）出る，(交渉に）入る」など，機能動詞的に働いて新聞の叙述を組み立てるものが多い。形容詞の「強い，高い」も，「力が強い」「背が高い」といった具体的な意味よりも，「反発が強い」「可能性が高い」といった抽象的な意味で新聞の叙述を組み立てることが多いと考

えられる。

以上，Ａ～Ｄ群にどのような語が所属しているか，そこに何らかの有意味な分離の傾向が見出せるかどうか検討し，寿岳(1967)による高頻度語の３分類のうち，Ａ群を中心に骨組み語が，Ｂ群・Ｃ群を中心に叙述語が，Ｄ群を中心にテーマ語が分布するという傾向を見出した。もう少し正確に言えば，骨組み語はＡ群を中心にＢ群にも，叙述語はＢ群・Ｃ群を中心にＡ群にも，テーマ語はＤ群を中心にＣ群にもというように分布し，概して，Ａ群からＤ群に向かって骨組み語→叙述語→テーマ語の順に，重なりつつも分離する傾向を見ることができる。骨組み語・叙述語・テーマ語が対数化特化係数散布度すなわち「語の使用量の層間変動の大きさ」によってこのように分離する理由については，以下のように考えることができる。

まず，骨組み語は，その語彙的意味が抽象的・形式的であり，それゆえに機能語的に，つまり，助詞・助動詞に準ずるように文や文章・談話のまさに骨組みを形作るように働くことから，どのような内容の文章・談話にも現れる確率が高くなり，したがって対数化特化係数散布度が小さくなる。一方，テーマ語は，その意味が具体的・実質的であり，さらに固有名詞のように個別的なものもあって，それゆえに文章・談話の内容を形作るものとして働くことから，特定の内容の文章・談話にのみ現れる確率が高くなり，したがって対数化特化係数散布度が大きくなる，というものである。寿岳(1967:26)も「骨組み語は，通時的共時的なさまざまな資料において，絶対的にあるいは比較的に動くことなく使われる語であるが，テーマ語は，その言語資料の内容によって動いてゆく語である」と指摘するが，この「動く／動かない」というとらえ方を，「使用範囲」のように「現れる／現れない」ではなく，文字通り「変動する／変動しない」と受け取れば，それは「何らかの基準による層別を施した語彙調査において，語の使用量の層間変動が大きい／小さい」と解釈することができ，「内容にもとづく層別においては，骨組み語の使用量は層間変動が小さくなり，テーマ語のそれは大きくなる」という結論が容易に得られることになる。

その上で，叙述語が骨組み語とテーマ語の中間に分布する理由は，まさにMcCarthy(1991=1995)が談話構成語を,機能語と内容語との中間にあって双方の性質を備えているような語と規定したように，叙述語も骨組み語とテーマ語と

の双方の性質を備えているから，つまり，骨組み語のように文や文章・談話の構成・構造を形作るものの，テーマ語のように実質的な語彙的意味をも有しているために，その構成・構造の作り方は骨組み語と必ずしも同じではなく，また，文章・談話の内容面とのかかわりも生じて，特定の内容の文章・談話に現れる確率も相対的に高くなるから，というようなことが考えられる。寿岳(1967:27)は，テーマ語と叙述語との違いは理論的にははっきり言い得ても，実際の判別には困難が伴うことがあるとして「はっきりした名詞というようなあらわれ方をするテーマのまわりをかこむ言葉，つまりさまざまの言葉のうち，話し手がいかに事象を捉えたかに関係あるような言葉はテーマ的か叙述的かあいまいになってくるのである」と述べるが，このような「テーマ的か叙述的かあいまい」な語こそが「叙述語」なのではないだろうか。

５．おわりに：高頻度語の「分離」から「分類」へ

以上，高頻度語の分離問題を計量語彙論の基本的な研究課題とする立場から，高頻度語を計量的に有効な精度で分離し得る新たな指標として「語の使用量の層間変動」に注目し，その大きさを「対数化特化係数散布度」によって測定する方法を提案して，現代新聞の語彙調査から得た高頻度上位約 100 語に適用した結果，対数化特化係数散布度の小さい方から概ね骨組み語→叙述語→テーマ語の順に分布（分離）するという傾向を見出し，その有効性を確認することができた。今後は，高頻度語の範囲を上位 100 語から 500 語，1000 語と増やしていって，同様の分離の傾向が観察されるか，確認していく必要がある。また，今回の８紙面という層別では，特定の紙面でその紙面特有の叙述を組み立てるような叙述語は，対数化特化係数散布度が大きくなってテーマ語の中に混じってしまい，まとまった形では取り出せない可能性が高い。こうした叙述語の発見には，また別の方法が必要になる。

とはいえ，高頻度語を対数化特化係数散布度によって観測することが，骨組み語・叙述語・テーマ語の分離につながるという見通しを得たことで，これら３分類についてのデータにもとづく質的な分析が進み，とりわけ骨組み語とテーマ語の中間に位置すると考えられる叙述語のより詳しい性質や機能が明らかになることが予想される。引き続き検討していきたい。

注

１）「この」が高校「物理」教科書で顕著に使用される理由については，石井(2019)の第１章および第２章（「低頻度語発生の文章機構(1)(2)」）を参照されたい。なお，本論文３節の表８に見るように，新聞では「この」の方が「その」よりも散布度が小さい。

２）正式名は“unidic-mecab-2.1.2_windows.exe”（https://ja.osdn.net/projects/unidic/，最終更新:2020-03-11 14:32)。現在，国立国語研究所コーパス開発センターから「現代書き言葉 UniDic」として公開されているものの旧バージョンである。

３）寿岳は「源氏物語の基礎語彙について一つの分類をほどこし」たとするが（寿岳 1967:18-19)，ここでいう「源氏物語の基礎語彙」とは論文第１節の表題にある通り「源氏物語においてよく使われた語」であり，寿岳の分類が（寿岳自身の意図とは異なるとしても）直接には「源氏物語の高頻度語」に対するものであることは明らかである。

引用文献

石井正彦(2015)「無性格語は実在するかー特化係数とその散布度による検討ー」斎藤倫明・石井正彦編『日本語語彙へのアプローチ：形態・統語・計量・歴史・対照』おうふう，pp.147-163

石井正彦(2019)『探索的コーパス言語学ーデータ主導の日本語研究・試論ー』大阪大学出版会

上田尚一(1981)『統計データの見方・使い方』朝倉書店

樺島忠夫(1980)「語彙」国語学会編『国語学大辞典』東京堂出版，pp.344-346

国立国語研究所(1964)『現代雑誌九十種の用語用字　第三分冊　分析』秀英出版

国立国語研究所(1984)『高校教科書の語彙調査Ⅱ』秀英出版（『国立国語研究所言語処理データ集６　中学校・高校教科書の語彙調査（フロッピー版)』秀英出版，1994）

寿岳章子(1967)「源氏物語基礎語彙の構成」『計量国語学』41，pp.18-32

高崎みどり(2012)「テクストの結束性に与る語彙とその機能について」『第２回

コーパス日本語学ワークショップ　予稿集』国立国語研究所言語資源系・コーパス開発センター，pp. 7–14

林　四郎(1971)「語彙調査と基本語彙」国立国語研究所報告 39『電子計算機による国語研究　Ⅲ』秀英出版，pp. 1–35

McCarthy, M. (1991) *Discourse Analysis for Language Teachers*, Cambridge University Press.（=1995, 安藤貞雄・加藤克美(訳)『語学教師のための談話分析』大修館書店）

付記

本稿は，2018～2021 年度日本学術振興会科学研究費補助金・基盤研究（C）（一般）「近現代日本語『基本語彙』史の記述に向けた新聞・和語動詞の『叙述語』化の研究」（研究代表者　金　愛蘭：課題番号 18K00612）による研究成果の一部である。

（文学研究科教授）