Discrimination of characteristic words in various fields

層別特徴語の判別

木村　繁

これは，報告31「電子計算機による国語研究」の林四郎論文34ページ「層による用語の特徴」（以下「層別特徴語」という）について，A0（41年版，朝日，朝刊の1月～6月分）に関して計算機で求めた一つの試行としての方法とその結果についての報告である。

A．調査資料

延べ218,231；異なり語形44,501のうち，各語の全体度数が7以上の長単位語（異なり3,067；延べ154,257；A0全体に占める比70.7%）について調べた。層別としては4種—すなわち文種別（G），位置別（P），署名態度別（S），話題別（T）—47層のうち，延べ語数の少ないG17漫画とP8，S10，T12広告 etc，（G14～17を集めたもの）を除く43層を調査対象とした。

同一の長単位でも層別が異なれば，あたかも異なるかのようにみなした〝層別語〟をここでは調査単位とする。従って調査単位数は，

3,067（異なり語）×43（層別）＝131,881（層別語）である。

なお，基本データは長単位A0の OUTPUT（〔表1〕参照）として作成済の層別度数台帳磁気テープを用いた。

B．判別の方法

層別語に関して，全体としてみた時の平均的な値（理論度数）を求め，実現度数が理論値と比べてなんらかの意味において著しくかけ離れているとき，その語を層別特徴語とし，それを次の操作によって決める。

1．理論度数は次式で計算される。

〔表　1〕　　Ａ０の述べと異なり

全体度数	異なり	累積異なり	累積延べ	比
＊500以上	30 語	30	79,909	36.6
400	5	35	82,110	37.6
300	13	48	86,538	39.7
200	22	70	92,068	42.2
100	71	141	102,270	46.8
80	40	181	105,790	48.5
60	67	248	110,393	50.5
40	133	381	116,782	53.5
30	162	543	122,352	56.2
20	381	924	131,464	60.3
15	361	1,285	137,537	63.0
10	716	2,001	145,908	66.8
9	271	2,272	148,347	67.8
8	345	2,617	151,107	69.2
7	450	3,067	154,257	70.7
6～1	41,434	44,501	218,231	100.0

＊度数順に｛
，(9542) の (9390) ・ (5145) を (4804)
に (4579) 「 (3841) 」 (3803) ¶ＭＯ (3791)
は (3653) が (3316) て (3244) と (2924)
。 (2483) た (2454) で (2290)
０ － 「 」 、る も １ から ある ない
こと し ２ ” いう この 歩

（理論度数）＝（全体度数）×（配分係数）

配分係数：層別延べ語数に対するＡＯ全延べ語数の割合

（値は〔表２〕参照）

すなわち，各語の全体度数を各層別の配分係数で比例配分した値が理論度数である。

2. 層別語の実現度数と理論度数を各々23クラスに分類し，〔図１〕のように行（Row）に理論度数を級区分し，列（Column）に実現度数を級区分した23×23の相関表を作る。ここで，実現度数が理論度数より大きい領域を〝正領域〟，小さい領域を〝負領域〟と定義する。〔図１〕では，左上角から45°の直線を引き，右上方の範囲が正領域，左下方の範囲が負領域となる。（級区分は〔表３〕を参照）

〔表　2〕

		層別延べ	配分係数 %	全体度数7以上の層別延べ	%	(全体度数に対する層別度数の回帰係数) B	A	(全体度数と層別度数の)相関係数 R
G 1	ニュース	44,850	(20.55)	32,889	(21.32)	0.2612	−2.4	0.9523
2	ニュース解説	7,401	(3.39)	5,530	(3.58)	0.0416	−0.3	0.9473
3	社説	7,962	(3.65)	5,952	(3.86)	0.0503	−0.6	0.9234
4	特集記事	7,986	(3.66)	5,906	(3.83)	0.0476	−0.5	0.9408
5	特別読物	9,256	(4.24)	6,936	(4.50)	0.0567	−0.6	0.9225
6	論評	4,085	(1.87)	3,124	(2.03)	0.0243	−0.2	0.9207
7	実用読物	5,769	(2.64)	4,184	(2.71)	0.0343	−0.4	0.9267
8	探訪	8,920	(4.09)	6,548	(4.24)	0.0480	−0.3	0.9513
9	ニュース展望	675	(0.31)	469	(0.30)	0.0040	−0.0	0.9287
10	通知	41,485	(19.01)	28,465	(18.45)	0.1484	1.8	0.7620
11	紹介	5,427	(2.49)	3,904	(2.53)	0.0322	−0.3	0.9445
12	読者	7,655	(3.51)	5,767	(3.74)	0.0418	−0.2	0.9381
13	コミニケ	4,415	(2.02)	3,310	(2.15)	0.0244	−0.1	0.9297
14	小説	1,769	(0.81)	1,397	(0.91)	0.0102	−0.1	0.8533
15	商業広告	24,958	(11.44)	15,224	(9.87)	0.0950	0.2	0.8992
16	〃 案内	35,558	(16.29)	24,618	(15.96)	0.0798	4.0	0.5836
17	〃 漫画	60	(0.03)					
計		218,231	(100.00)	154,257	(100.00)			

P 1	見出し	6,746	(3.09)	4,357	(2.82)	0.0300	−0.1	0.9434
2	標題下	994	(0.46)	681	(0.44)	0.0053	−0.0	0.9330
3	リード	3,629	(1.66)	2,679	(1.74)	0.0222	−0.2	0.9385
4	本文	110,704	(50.73)	81,981	(53.15)	0.6562	−6.3	0.9440
5	情報源	1,325	(0.61)	960	(0.62)	0.0061	0.0	0.7890
6	表	30,833	(14.13)	21,279	(13.79)	0.0871	2.6	0.5059
7	図.表.写真説明	1,655	(0.76)	1,047	(0.68)	0.0079	−0.1	0.9307
8	G 14 ～ G 17	62,345	(28.57)					
計		218,231	(100.00)	154,257	(100.00)			

S 1	無署名	94,657	(43.37)	67,961	(44.06)	0.4814	−2.1	0.9873
2	信社	7,070	(3.24)	5,176	(3.36)	0.0408	−0.4	0.9399
3	冒頭	8,449	(3.87)	6,424	(4.16)	0.0493	−0.4	0.9231
4	末尾(外部)	9,732	(4.46)	7,279	(4.72)	0.0549	−0.4	0.9572
5	末尾(記者)	11,033	(5.06)	8,076	(5.24)	0.0586	−0.3	0.9660
6	末尾(略称)	4,210	(1.93)	3,154	(2.04)	0.0239	−0.2	0.9325
7	外電冒頭	6,811	(3.12)	5,039	(3.27)	0.0401	−0.4	0.9337
8	社を代表	5,868	(2.69)	4,434	(2.87)	0.0374	−0.4	0.9173
9	無署名(外部)	8,056	(3.69)	5,441	(3.53)	0.0283	0.3	0.7907
10	広告	62,345	(28.57)					
計		218,231	(100.00)	154,257	(100.00)			

T 1	政治	15,253	(6.99)	11,408	(7.40)	0.0913	−0.9	0.9347
2	外交	2,962	(1.36)	2,233	(1.45)	0.0176	−0.2	0.9293
3	経済	23,814	(10.91)	16,378	(10.62)	0.0911	0.8	0.7836
4	労働	2,666	(1.22)	1,984	(1.29)	0.0160	−0.2	0.9331
5	社会	26,263	(12.03)	19,123	(12.40)	0.1532	−1.5	0.9431
6	国際	16,725	(7.66)	12,464	(8.08)	0.0986	−0.9	0.9416
7	文化	16,303	(7.47)	12,117	(7.86)	0.0932	−0.7	0.9514
8	地方	5,417	(2.48)	3,809	(2.47)	0.0273	−0.1	0.9617
9	スポーツ	20,536	(9.41)	14,992	(9.72)	0.0972	−0.0	0.8676
10	婦人	9,957	(4.56)	7,335	(4.76)	0.0590	−0.6	0.9198
11	芸能	15,990	(7.33)	11,141	(7.22)	0.0702	0.1	0.8292
12	広告	62,345	(28.57)					
計		218,231	(100.00)	154,257	(100.00)			

〔表　3〕

			理論値（i）				実数値（j）			
級番号	級の（下）限界	＊中央値	層別語	平均 Mi	標準偏差 Si	Mi＋2Siを越える級番号	変異係数 Vi＝100×Si/Mi	層別語	Mj	Sj	Mj＋2Sjを越える級番号	Vj
0#	0	0	61,204	0.2	0.7	3#(3)	350	89,058	0.9	3.0	8#(8)	330
1#	1	1	38,764	0.8	1.7	5#(5)	210	16,146	1.3	3.4	9#(9)	260
2#	2	2	11,162	1.9	3.3	10#(10)	170	7,234	1.9	5.0	11#(15)	260
3#	3	3	5,441	2.9	4.2	11#(15)	140	4,064	2.5	4.4	11#(15)	180
4#	4	4	3,620	4.0	5.3	11#(15)	130	2,802	3.2	6.2	11#(15)	190
5#	5	5	2,267	4.7	5.9	12#(20)	130	1,981	3.8	7.1	12#(20)	190
6#	6	6	1,579	5.6	7.5	13#(30)	130	1,599	4.1	6.2	12#(20)	150
7#	7	7	1,190	6.9	8.0	13#(30)	120	1,367	4.5	8.1	12#(20)	180
8#	8	8	778	6.8	8.1	13#(30)	120	1,006	5.2	6.3	12#(20)	120
9#	9	9	612	7.3	8.4	13#(30)	120	736	6.0	6.6	12#(20)	110
10#	10	12	1,963	11.1	13.1	14#(40)	120	2,151	8.9	14.8	14#(40)	170
11#	15	17	829	16.4	19.9	15#(60)	120	985	13.3	25.7	16#(80)	190
12#	20	24	778	23.8	25.9	16#(80)	110	941	19.4	40.0	17#(100)	210
13#	30	34	383	34.0	39.1	17#(100)	110	418	29.2	37.1	17#(100)	130
14#	40	49	356	49.4	49.8	18#(150)	100	408	48.6	86.1	19#(200)	180
15#	60	69	205	72.5	60.1	19#(200)	83	224	52.7	68.9	19#(200)	130
16#	80	89	128	94.1	68.2	20#(300)	72	98	85.4	95.1	20#(300)	110
17#	100	124	207	124.0	121.5	21#(400)	98	188	92.9	95.3	20#(300)	100
18#	150	174	112	191.3	139.2	22#(500)	73	115	158.5	137.5	21#(400)	87
19#	200	249	92	296.5	188.3	……	63	134	195.5	153.0	22#(500)	78
20#	300	349	66	492.8	259.5	……	53	57	335.2	201.5	……	60
21#	400	449	34	472.9	326.1	……	69	36	362.8	192.1	……	53
22#	500	800	111	645.8	258.4	……	40	133	618.8	230.3	……	37
		（計）	131,881				（計）	131,881				

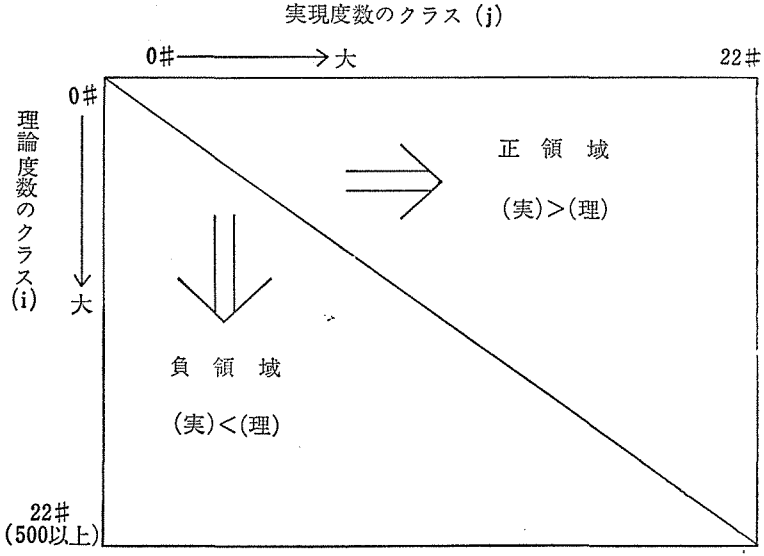
3. 累積千分比の計算

正領域では，理論値のクラス毎に実現値の大きい方から，すなわち〔図1〕では右側から横の方向に，累積千分比を計算する。負領域では，実現値のクラス毎に理論値の大きい方から，すなわち〔図1〕では下側から上の方向に，累積千分比を計算する。

4. 実現値あるいは理論値のクラスを固定―すなわち，相関表を縦割りあるいは横割り―した時のクラス別の平均M，標準偏差Sを求める。

（〔表3〕を参照）

5. （M＋2S）を越えるクラスに属す層別語を特徴語とした。このとき，正領域あるいは負領域に属すかの2通りの意味での層別特徴語が存在する。そこで，判別テーブルとしては，層別特徴語に属する領域（層別特徴領域）については累積千分比別に判別マークを与える。



〔図　1〕

正領域では，

＋（0.1%以下）

A（1.0%以下）

B（2.0%以下）

C（2.0%を越える層別特徴領域）

負領域では，

－（0.1%以下）

J（1.0%以下）

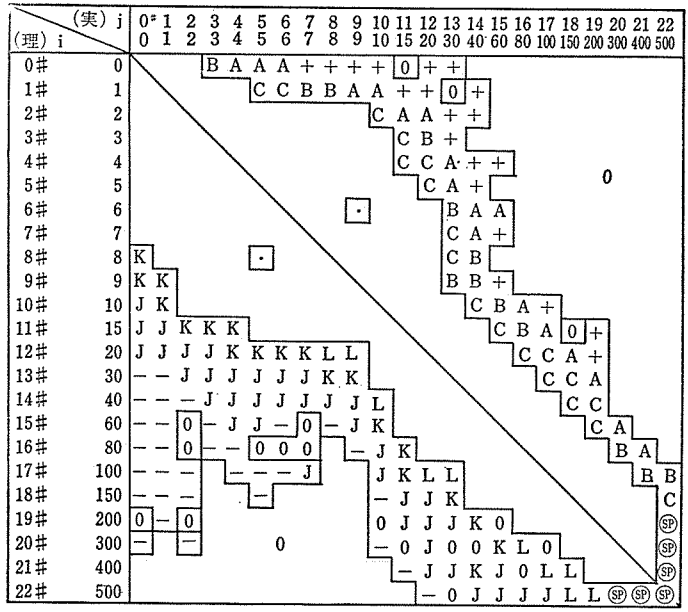
K（2.0%以下）

L（2.0%を越える層別特徴領域）

の記号を与え，層別特徴領域に属さない（M＋２Ｓ）より小さい範囲では，・を与える。（〔表４〕参照）

※なお，〔表４〕では，層別語が存在しないマス

〔表　4〕



目には0を与えている。また，クラス番号22＃（度数500以上）の部については，Ⓢ を与え，別に特徴語を判別することにした。

C．判別の吟味

1. 上のBの5．の判別基準では，理論度数0（小数第一位四捨五入して）の時，実現度数3以上を層別特徴語としている。これは，配分係数7.2％以下，全体度数7の場合，層別度数が3以上である時，正領域での特徴語となる。また，理論度数8以上で，実理度数0の時，負領域での特徴語となる。
2. （M＋2S）で層別特徴語の境界線を引くと，級番号が大きくなる時，境界線が逆行しない（すなわち，増加函数）。ただし，千分比による等比線は逆行している。

3. 全体度数X，層別度数Yとして，最小自乗法により，回帰直線Y＝BX＋A及び，X，Yの相関係数Rを求めた。（〔表２〕参照）

相関係数0.9以下の層別としては，次のものがある。

0.58＝（G16案内広告）＜（G10通知）＜（G14小説）＜（G15商業広告）

0.51＝（P７図，表，写真の説明）＜（P５情報源）

0.79＝（S９無署名（外部））

0.78＝（T３経済）＜（T11芸能）＜（T９スポーツ）

これらの層では，層別特徴語が多いと考えてよいだろう。

D．結果及び分析

1. OUTPUT としては，度数順に各層の特徴語判別マークをラインプリントした。長単位語は，度数順語彙表から人手により転写した。〔なお，作表の仕方としては，ＩＲ 的見地から必要な項目だけを OUTPUT することも考えられるが，漢テレに負荷がかからないようにするため，これを基本表とした。〕

（例）

長単位・位置	FREQ	(G)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	(P)～(S)～(T)
た 1P14L	2454	16	㉕	・	・	・		・	・	・	・	L	・	・	・	・	・	J	
いる 20L	1007	15	・	・	・	・	・	・	・	・	・	・	・	・	・	・	L	－	
ある 24L	776	15	・	・	・	・	・	・	・	・	・	L	・	・	・	・	・	L	（以下略）
ない 25L	759	15	・	・	・	・	・	・	・	・	・	L	・	・	・	・	・	－	
ます 2P12L	386	15	J	K	K	K	・	・	C	K	・	J	・	・	B	・	・	・	
だ 3P04L	271	16	・	・	・	・	・	C	・	・	・	・	・	・	・	・	・	J	
です 24L	191	14	J	・	・	・	・	・	・	・	・	・	・	・	C	・	・	・	

(G). (P). (S). (T)には０でない層の数をプリント。

※ この表は，４層別を除く全層別の一種の濃淡を示しているとも考えられる。

2. この表から各層の正領域の特徴語を集めた。

Ｇ１ニュースについて層別特徴語を列記する。

G1・ニュース
二十七日，二十二日，十三日，二週間，日本時間
ジャカルタ，ベトナム，アラブ連合，カイロ，北京，マレーシア
桑田，重雄さん，三井物産，日銀，公明，公明党，農林省，非同盟，発，ロイター，UPI
委員会，委員長，付近，教会，史跡，四条件，書簡，上空，同党，晴着，両党
会談，デモ，開催，協定，再編成，支持，成立，提案，北爆，満足，寝，語っ，出席し，かけつけ
新たな

3. 人称代名詞について特徴語となる層別を記す。

4. この他に例えば

1) 社説では，助詞，助動詞としては「なら，うと，なければ，べき，かは」が特徴語になっいる。

2) 経済，スポーツなどは，数字が多い。

など，語種，品詞などを付加情報に与えて，整理してみるとおもしろ

長単位	全体度数	順位（）内は記号を除く順位	正領域（実）＞（理）	負領域（実）＜（理）
私	118	125 (106)	G12読者，G13コミニケ，S3冒頭，S4末尾（外部），S6（略称），T10婦人	G16案内広告 T3経済，T9スポーツ
あなた	55	267 (242)	G13 コミニケ T10 婦人	G1 ニュース
われわれ	34	460 (429)	G3 社説 S3 冒頭，S8 社を代表	————
彼	28	573 (542)	G11 紹介，G13 コミニケ S3 冒頭，T7 文化	————
わが	23	733 (696)	————	————
私たち	20	871 (834)	G5 特読，S4末尾(外部) T7 文化	————
彼女	10	1803 (1758)	G5 特読	————
ぼく	9	2002 (1957)	G5特読，G12読者，S3冒頭，S4末尾(外部)，T8地方	————
わたし	9	2002 (1957)	G14 小説	————

いと思う。

3) この分析は長単位についてであるが，「経験者」という語がG16案内広告において特徴語であるというような，β単位とはまた別の結果が得られる。その他，余談だが，G16案内広告に「25歳」が特徴語となっているが，これはこの年令が転職あるいは求職が多いことを示していないだろうか。