

国立国語研究所学術情報リポジトリ

Key-words for automatic abstracting of literary texts

メタデータ	言語: jpn 出版者: 公開日: 2017-03-31 キーワード (Ja): キーワード (En): 作成者: 田中, 章夫, TANAKA, Akio メールアドレス: 所属:
URL	https://doi.org/10.15084/00001023

自動抄録処理における

キー・ワードの性格

田 中 章 夫

0・はじめに

自動抄録の理論とその処理過程における諸問題は、言語研究、特に計量言語学の諸研究とかなり密接な関係があり、相互に影響しあうところが、きわめて多い。しかし、今まで、この問題についての言語研究の側からの発言は、あまり見られなかった。それは一つには、現在までの自動抄録が、科学技術論文など、主として、特定な分野に限られた専門的なものの処理を旨として進んできたことによる。特定な専門分野のものであれば、おおよそのところがわかれば一応の役に立つといった性格が強いためか、原文の性質や抄出文の質などについては、立ちいった検討がなされなかったようである。また、科学技術論文のように、情報密度の高い論理的な文章を主に扱ってきたため、比較的単純な計量的処理だけにたよっていても、ある程度の成果をあげたことも、言語的な方面の研究とつながらなかつた一因ではあるまいか。

しかし、一般の文章を対象として、抄録処理を進めるとなると、さまざまな性格の文章が現われてくるうえに、抄出された文章の、文章としての自然さやまとまりが問題にされ、すくなくとも、従来のように、単純な計量的操作だけではカバーしきれないように思われる。

自動抄録の基本的な操作は、対象とする原文の中から、その文章の記述内容をなっているキー・ワードを選定し、それを手がかりに一定量のセンテンスを抽出して抄録を作りあげることであるが、これは、いわば文章構成の一種のシミュレートにほかならない。そして、このプロセスの中には、文章に使われている各単語の性格の検討をはじめ、文章構成の理論など、各所に言語学的な研究課題が提供されている。

今回は、一般的な文章の一例として文学作品をとりあげ、その抄録の実験を試みるとともに、キー・ワードとなるべき単語の統計的あるいは意味的な性格と、それによって抽出されてくるセンテンスの量の問題とを中心に考えてみようと思う。

1. キー・ワード密度による自動抄録法

かつて、水谷静夫氏は、論文「統計的自動抄録法の問題点」^(注1)において、P・H・ルーンやV・A・オスワルドらによる自動抄録理論に検討を加えるとともに、種々の独創的な考え方と手法を示した。水谷氏は、また、論文「抄録を作る機械」^(注2)において、雑誌「中央公論」の論説を材料にした実験を試みているが、P・H・ルーンの方法も、水谷氏のこの実験も、基本的には、つぎの二つの前提に立っている。

- 抄録を作るべき文献でポイントとなる事項に関係の深い言葉は、概して、その文献に繰り返し出現する（ただし、逆は必ずしも真ではない）。
- そのような言葉を一文中に数多く含む文は、その文献で、概して肝腎なところである（したがって、それに該当する文を選出して並べれば、抄録の一応の目的は達せられるであろう）。

こうした前提に立って、P・H・ルーンは、つぎのような自動抄録法を提唱した。^(注3)

- ①抄録しようとする文献の本文を計算機に入力し、その本文の語彙調査をする。
- ②繰り返し出現する、すなわち、よく使われている語の中から、どの文章にもよく出現している語を除いて、残りを意味語 (a Set of Significant Words) とする。
- ③その意味語を一文中に含む割合の大小によって、各文に階級をつける。
- ④あらかじめ定めてある規準によって、各文をそれぞれ抄録中に採用するかどうかを判定する。
- ⑤上で採用が決定した文を、もとの文献に出現した順に並べて抄録とする。

—HITAC APPLICATION「電子計算機による情報検索」より—

以上述べた P・H・ルーンの系統の自動抄録法の、最も大きな特徴は、上記の③にあるように、1 センテンス内のキー・ワード（ルーンの言う a Set of Significant Words）の割合の大小によって、文の抽出を行なっていくところにある。このような、1 センテンス内のキー・ワードの密度によって抄録する方法は、すでに水谷氏も指摘しているように、

「文体に一貫性がない場合」や「構文というか文のまとまりと、内容のまとまりが、うまく一致しない場合」にはキー・ワード密度は、あまり有効に働かないおそれがある。^(注4) こうした点は、情報密度が高く、論理的な科学技術論文などの文章では、あまり問題にならないのかもしれないが、いくつか実験を試みたところでは、文学作品のような一般の文章の場合には、つぎの二点において、キー・ワード密度方式は、致命的であった。

その第一は、一般の文章、特に小説などにおいては、キー・ワード密度の高い文が、対話部分に偏在しやすいことである。小説類の場合、キー・ワードとして重要な役割を果たすものは、主人公やヒロインの名前であるが、キー・ワードが、この種のものであると、キー・ワード密度が最も高くなりやすいのは「誰々さん！」といったような、単なる呼びかけのセンテンスということになってしまう。

第二は、「A氏はB子に話しかけた」という類の、場面説明に過ぎないセンテンスが、やはり、キー・ワード密度の高い文として、数多く抽出されて来ってしまう点である。

上に述べた第一の点は、結局、文章の中に会話の文体が混在しているために生じるものであり、水谷氏の言う「文体に一貫性がない場合」に当る。また第二の点は、場面の説明というような文章の主題の上からみれば、あまり重要でないセンテンスにキー・ワードが集まりすぎるためであり、これは結局、筋の展開すなわち内容のまとまりと文のまとまりとが、うまく一致していない場合ということになる。

最後にもう一つ、キー・ワード密度方式の都合のわるい点を挙げると、科学技術の論文など、論理が一貫している文章では、さほど問題にならないかもしれないが、一般の文章、特に小説などにおいては、話題の変換や挿入、あるいは

は場面の転換などが、かなり自在に行なわれるという点である。そのためにキー・ワード密度だけで追っていたのでは、話の筋の展開が、なかなか、うまくとらえられないことが多い。つまりキー・ワード方式は、話の筋の変化に弱いというわけである。もちろん、そうした場合、段落ごとに処理するか、パラグラフごとに抄出するかといった方法によって、この欠点を、ある程度カバーすることもできようが、論文などと違って、本来、文章の筋道や論理の一貫性を、必ずしも重視しない一般の文章の場合には、キー・ワード方式一本で抄録を進めていくことには、かなり無理があるのではないかと思うのである。

水谷氏も、論文「統計的自動抄録法の問題点」^(注1)において、章節ごとに抄録を進める試みを発表しているが、これも一つには、頭からキー・ワード密度方式のみで進めたのでは、章節ごとの論旨の展開が、うまくとらえられないためである。

以上のような理由によって、今回の自動抄録の実験においては、キー・ワード密度の大小によって、文を抽出する方式は採らなかった。

注1) 「計量国語学・27号」所収

注2) 「言語生活・137号」所収

注3) Luhn P.H. (1958)“The Automatic Creation of Literature Abstract”
(IBM Journal・vol. 2)

注4) 「計量国語学・27号」8ページ

2. キー・ワードについての仮説

自動抄録処理を行なう場合、そのキー・ワードというものは、対象となる文章の、筋の展開の上で重要なことばであって、その文章の主題に密接な関係があり、かつ文章の内容をうまく代表するものであってほしい。簡単にいえば、その文章を特徴づける単語、いかにもその文章らしい単語を仮定しているわけである。

もし、そういう単語が、キー・ワードとして適切に選ばれているならば、その語が、文章の中で最初に出現するセンテンスは、話の切り出しや話題の転換あるいは場面の設定や変換が行なわれる個所であり、話の筋の展開の上で重要なところであるはずである。これが、ここに発表する自動抄録法の第一の仮説

表 1 諸作品の高頻度語

寒山拾得 (森鷗外)											
延べ語数 1907											
異り語数 765											
順位	見出し	度数	使用率	累積 度数	累積 使用率						
			%		%						
1	云ふ	61	3.20	61	3.20			27	2.54	151	14.22
2	ゐる	58	3.04	119	6.24			6	1.88	171	16.10
3	する	55	2.88	174	9.12			6	1.88	191	17.98
4	聞	42	2.20	216	11.33			8	1.60	208	19.59
5	こと(事)	32	1.68	248	13.00			9	1.51	224	21.09
6	ある	23	1.21	271	14.21			10	1.22	237	22.32
6	それ	23	1.21	294	15.42		◎	10	1.22	250	23.54
6	もの	23	1.21	317	16.62			12	1.13	262	24.67
9	なる	20	1.05	337	17.67			12	1.13	274	25.80
9	見る	20	1.05	357	18.72		◎	14	1.04	285	26.84
11	をる	19	1.00	376	19.72			14	1.04	296	27.87
12	僧	18	0.94	394	20.66			16	0.94	306	28.81
12	人	18	0.94	412	21.60			16	0.94	316	29.76
14	道	17	0.89	429	22.50			16	0.94	326	30.70
15	無い	16	0.84	445	23.34			16	0.94	336	31.64
16	これ	14	0.73	459	24.07			9	0.85	345	32.49
17	出る	13	0.68	472	24.75			21	0.75	353	33.24
17	水	13	0.68	485	25.43			21	0.75	361	33.99
19	来る	12	0.63	497	26.06			21	0.75	369	34.75
19	行く	12	0.63	509	26.69			21	0.75	377	35.50
21	其	11	0.58	520	27.27			21	0.66	384	36.16
21	台州	11	0.58	531	27.84		○	21	0.66	391	36.82
21	道翹	11	0.58	542	28.42			21	0.66	398	37.48
21	時	11	0.58	553	29.00						
25	寒山	10	0.52	563	29.52						
25	豊干	10	0.52	573	30.05						
清兵衛と瓢箪 (志賀直哉)						小僧の神様 (志賀直哉)					
延べ語数 1062						延べ語数 1895					
異り語数 429						異り語数 648					
順位	見出し	度数	使用率	累積 度数	累積 使用率	順位	見出し	度数	使用率	累積 度数	累積 使用率
			%		%				%		%
1	清兵衛	33	3.11	33	3.11	1	いる	100	5.28	100	5.28
2	いる	32	3.01	65	6.12	2	それ	64	3.38	164	8.65
3	瓢箪	31	2.92	96	9.04	3	する	60	3.17	224	11.82
4	彼	28	2.64	124	11.68	4	其	50	2.64	274	14.46
						5	言う	39	2.06	313	16.52
						6	事	36	1.90	349	18.42
						7	彼	30	1.58	379	20.00
						7	さう	30	1.58	409	21.58
						9	或	29	1.53	438	23.11
						10	小僧	27	1.42	465	24.54
						11	もの	24	1.27	489	25.80
						12	仙吉	23	1.21	512	27.02

13	来る		22	1.16	534	28.18	18	又		8	0.67	355	29.91
13	何		22	1.16	556	29.34	23	あげる		7	0.59	362	30.50
13	前		22	1.16	578	30.50	23	教員	○	7	0.59	369	31.09
16	行く		21	1.11	599	31.61	23	しまう		7	0.59	376	31.68
17	鮪屋	○	20	1.06	619	32.66	23	手	○	7	0.59	383	32.27
17	なる		20	1.06	639	33.72	23	ない		7	0.59	390	32.86
19	自分		19	1.00	658	34.72	密柑 (芥川竜之介) 延べ語数 1020 異り語数 552						
20	思う		18	0.95	676	35.67							
21	客	○	17	0.90	693	36.57	順位	見出し	度数	使用率	累積 度数	累積 使用率	
22	お		16	0.84	709	37.41				%		%	
22	ない		16	0.84	725	38.26	1	ゐる		30	2.94	30	2.94
24	知る		15	0.79	740	39.05	2	私		29	2.84	59	5.78
24	から		15	0.79	755	39.84	3	する		28	2.75	87	8.53
26	考へる		14	0.74	769	40.58	4	小娘	◎	16	1.75	103	10.10
26	やう		14	0.74	783	41.32	5	ある		15	1.47	118	11.57
真 鶴 (志賀直哉) 延べ語数 1187 異り語数 536							5	その		15	1.47	133	13.04
順位	見出し	度数	使用率	累積 度数	累積 使用率		7	この		13	1.27	146	14.31
			%		%		8	汽車	◎	11	1.08	157	15.39
1	彼	56	4.72	56	4.72		9	窓	◎	9	0.88	166	16.27
2	いる	41	3.45	97	8.17		9	中		9	0.88	175	17.16
3	する	38	3.20	135	11.37		11	見る		8	0.78	183	17.94
4	いう	19	1.60	154	12.97		11	くる		8	0.78	191	18.73
4	女	◎	19	1.60	173	14.57	11	言ふ		8	0.78	199	19.51
6	弟	◎	17	1.43	190	16.01	11	そう		8	0.78	207	20.29
7	その		15	1.26	205	17.27	15	トンネル	◎	7	0.69	214	20.98
8	来る		14	1.18	219	18.45	15	思ふ		7	0.69	221	21.67
8	こと		14	1.18	233	19.63	15	なる		7	0.69	228	22.35
10	それ		11	0.93	244	20.56	15	それ		7	0.69	235	23.04
10	行く		11	0.93	255	21.48	19	出す		6	0.59	241	23.63
10	なる		11	0.93	266	22.41	19	ない		6	0.59	247	24.22
13	おもう		10	0.84	276	23.25	19	戸	◎	6	0.59	253	24.80
13	自分		10	0.84	286	24.09	19	何		6	0.59	259	25.39
13	そして		10	0.84	296	24.94	19	踏切	○	6	0.59	265	25.98
13	見る		10	0.84	306	25.78	羅生門 (芥川竜之介) 延べ語数 1852 異り語数 678						
17	気		9	0.76	315	26.54	順位	見出し	度数	使用率	累積 度数	累積 使用率	
18	今		8	0.67	323	27.21				%		%	
18	この		8	0.67	331	27.89	1	する	74	4.00	74	4.00	
18	水兵	○	8	0.67	339	28.56	2	いる	65	3.51	139	7.51	
18	何		8	0.67	347	29.23							

3	ある		54	2.92	193	10.42	10	絵	○	19	1.78	320	29.96
4	下人	◎	44	2.38	237	12.80	10	氏		19	1.78	339	31.74
5	この		39	2.11	276	14.90	12	この		18	1.69	357	33.43
6	その		32	1.73	308	16.63	13	様(ヨウ)		16	1.50	373	34.93
7	事		31	1.67	339	18.30	14	前		12	1.12	385	36.05
8	いふ		29	1.57	368	19.87	15	A	○	11	1.03	396	37.08
9	老婆	◎	28	1.51	396	21.38	15	もの		11	1.03	407	38.11
10	ない		27	1.46	423	22.84	17	見る		9	0.84	416	38.95
11	やう		25	1.35	448	24.19	18	そう		8	0.75	424	39.70
12	上		23	1.24	471	25.43	18	数年		8	0.75	432	40.45
13	一		19	1.03	490	26.46	18	来る		8	0.75	440	41.20
13	それ		19	1.03	509	27.48	21	見せる		7	0.66	447	41.85
15	なる		17	0.92	526	28.40	22	見える		6	0.56	453	42.42
16	門	○	16	0.86	542	29.27	22	作品	○	6	0.56	459	42.98
16	見る		16	0.86	558	30.13	22	中		6	0.56	465	43.54
18	さう		15	0.81	573	30.94	22	行く		6	0.56	471	44.10
19	死骸	○	14	0.76	587	31.70	22	寄る	○	6	0.56	477	44.66
20	何		13	0.70	600	32.40	刺青(谷崎潤一郎) 延べ語数 1531 異り語数 868						
21	男	○	11	0.59	611	32.99							
21	来る		11	0.59	622	33.59							
23	出す		10	0.54	632	34.13							
23	時	○	10	0.54	642	34.67	順位	見出し	度数	使用率	累積 度数	累積 使用率	
23	髪		10	0.54	652	35.21	1	居る		33	2.16%	33	2.16%
23	上る		10	0.54	662	35.75	2	娘	◎	28	1.83	61	3.98
27	雨	○	9	0.49	671	36.23	3	ある		27	1.76	88	5.75
27	中		9	0.49	680	36.72	3	する		27	1.76	115	7.51
27	梯子	○	9	0.49	689	37.20	5	女	◎	23	1.50	138	9.01
窓(堀辰雄) 延べ語数 1068 異り語数 396							5	言う		23	1.50	161	10.52
順位 見出し 度数 使用率 累積 度数 累積 使用率							7	清吉	◎	22	1.44	183	11.95
							7	その		22	1.44	205	13.39
順位 見出し 度数 使用率 累積 度数 累積 使用率							9	彼		19	1.24	224	14.63
							9	お前		19	1.24	243	15.87
順位 見出し 度数 使用率 累積 度数 累積 使用率							11	顔	◎	14	0.91	257	16.79
							12	刺青	◎	13	0.85	270	17.64
順位 見出し 度数 使用率 累積 度数 累積 使用率							12	この		13	0.85	283	18.48
							14	見る		12	0.78	295	19.27
順位 見出し 度数 使用率 累積 度数 累積 使用率							14	男	◎	12	0.78	307	20.05
							14	人		12	0.78	319	20.84
順位 見出し 度数 使用率 累積 度数 累積 使用率							17	私		11	0.72	330	21.55
							18	それ		10	0.65	340	22.21
順位 見出し 度数 使用率 累積 度数 累積 使用率							18	こう		10	0.65	350	22.86
							18	こう		10	0.65	350	22.86

20	なう		9	0.59	359	23.45	24	らしい		10	0.81	480	38.71
20	絵	◎	9	0.59	368	24.04	27	込む		9	0.73	489	39.44
22	心	◎	8	0.52	376	24.56	28	日		8	0.65	497	40.08
23	行く		7	0.46	383	25.02	28	障子	○	8	0.65	505	40.73
23	こと		7	0.46	390	25.47	28	上		8	0.65	513	41.37
23	針	○	7	0.46	397	25.93	山椒魚（井伏鱒二） 延べ語数 1345 異り語数 552						
23	中		7	0.46	404	26.39							
23	今		7	0.46	411	26.85							
23	もう		7	0.46	418	27.30							
29	親方	○	6	0.39	424	27.69	順位	見出し	度数	使用率	累積 度数	累積 使用率	
29	お前		6	0.39	430	28.09				%	%		
ジガ峰（島木健作） 延べ語数 1240 異り語数 379							1	する	58	4.31	58	4.31	
順位	見出し	度数	使用率	累積 度数	累積 使用率		2	ある	55	4.09	113	8.40	
			%	%	%		3	彼	45	3.35	158	11.75	
1	ある	40	3.23	40	3.23		4	こと	30	2.23	188	13.98	
2	いる	39	3.15	79	6.37		5	山椒魚	29	2.16	217	16.13	
3	する	33	2.66	112	9.03		6	いる	28	2.08	245	18.22	
4	私	32	2.58	144	11.61		7	ない	23	1.71	268	19.93	
5	しる	25	2.02	169	13.63		8	岩屋	20	1.49	288	21.41	
6	穴	24	1.94	193	15.56		9	等	15	1.12	303	22.53	
7	ない	23	1.85	216	17.42		10	一	14	1.04	317	23.57	
8	一	20	1.61	236	19.03		11	もの	12	0.89	329	24.46	
9	こと	19	1.53	255	20.56		11	言う	12	0.89	341	25.35	
10	よう	18	1.45	273	22.02		11	出る	12	0.89	353	26.25	
11	彼等	17	1.37	290	23.39		14	水面	11	0.82	364	27.06	
11	何	17	1.37	307	24.76		15	お前	9	0.67	373	27.73	
11	見る	17	1.37	324	26.13		15	しまう	9	0.67	382	28.40	
14	それ	16	1.29	340	27.42		15	行く	9	0.67	391	29.07	
15	いう	15	1.21	355	28.63		18	自分	8	0.59	399	29.67	
16	時	13	1.05	368	29.68		18	外	8	0.59	407	30.26	
16	その	13	1.05	381	30.73		18	そこ	8	0.59	415	30.86	
18	ジガ峰	12	0.97	393	31.69		18	その	8	0.59	423	31.45	
18	また	12	0.97	405	32.66		18	見る	8	0.59	431	32.04	
18	思う	12	0.97	417	33.63		18	なか	8	0.59	439	32.64	
21	飛ぶ	11	0.89	428	34.52		18	なる	8	0.59	447	33.23	
21	行く	11	0.89	439	35.40		18	何	8	0.59	455	33.83	
21	もの	11	0.89	450	36.26	一房の葡萄（有島武郎） 延べ語数 1882 異り語数 627							
24	間	10	0.81	460	37.10	順位	見出し	度数	使用率	累積 度数	累積 使用率		
24	なる	10	0.81	470	37.90				%	%			
						1	僕	101	5.37	101	5.37		

				山月記(中島敦)						
				延べ語数		1878				
				異り語数		842				
				順位	見出し	度数	使用率	累積 度数	累積 使用率	
							%		%	
2	する	60	3.19	161	8.55					
3	いる	58	3.08	219	11.64					
4	先生	◎	34	1.81	253	13.44				
5	なる		30	1.59	283	15.04				
6	こと		27	1.43	310	16.47				
7	その		25	1.33	335	17.80	1	する	57 3.04	
8	いう		23	1.22	358	19.02	2	己	41 2.18	
9	ジム	◎	22	1.17	380	20.19	3	こと	31 1.65	
9	見る		22	1.17	402	21.36	4	ない	30 1.60	
11	思う		21	1.12	423	22.48	5	いる	28 1.49	
12	しまう		20	1.06	443	23.54	6	なる	27 1.44	
13	絵具	◎	19	1.01	462	24.55	6	ある	27 1.44	
14	顔	◎	16	0.85	478	25.40	8	自分	24 1.28	
15	行く		15	0.80	493	26.20	9	声	◎	22 1.17
15	ある		15	0.80	508	26.99	10	言う		21 1.12
17	中(なか)		14	0.74	522	27.74	11	今		21 1.12
18	いい		13	0.69	535	28.43	12	李徴	◎	20 1.06
18	もう		13	0.69	548	29.12	13	袁倬	◎	19 1.01
20	はいる		12	0.64	560	29.76	14	人間	◎	18 0.96
21	手	○	11	0.58	571	30.34	15	中		17 0.91
21	達		11	0.58	582	30.92	16	その		16 0.85
21	葡萄	○	11	0.58	593	31.51	17	この		15 0.80
24	来る		10	0.53	603	32.04	17	虎	◎	15 0.80
24	一		10	0.53	613	32.57	19	時		14 0.75
24	色	○	10	0.53	623	33.10	20	それ		12 0.64
24	なに		10	0.53	633	33.63	20	濃	○	12 0.64
							20	しかし		12 0.64
							20	見る		12 0.64
							24	一		11 0.59
							24	わかる		11 0.59
							26	これ		10 0.53
							26	誰		10 0.53
							26	もの		10 0.53

である。この仮説は、すなわち、キー・ワード初出センテンスの重視ということにほかならない。

一般に、このようなキー・ワードすなわち「いかにも、その文章らしい単語」というものは、その文章の用語について、頻度調査を行なうと、頻度順位の比較的上位のところに並んでくる。いま、いくつかの短篇について、用語の頻度調査の結果を掲げると表1の通りである。

どの作品の場合も、話題の主である主人公や登場人物の名まえ、あるいは、話の背景として重要な場面や道具立てに関することばは、ほぼ上位20位から30位ぐらいまでのところに、一応、現われてくる。したがって、P・H・ルーンが、すでに指摘しているように、これら高頻度の単語群の中から、「する」とか「いる」とか「とき」とかいった類の、どんな文章にも必ず出てくるような、ありふれたことば、別な言い方をすれば、その文章を特徴づけない「無性格な単語」をふるい落せば、一応「その文章らしい単語（表1の◎印・○印のような語）すなわちキー・ワード」が選び出されてくることになる。このように、対象とする文章にとって重要な単語は、高い頻度で現われてくるというのが、第二の仮説である。ただし、実際の操作においては、頻度順位を目やすにすると、文章のボリュームが影響するので、累積使用率を採用することにした。たとえば、頻度順位の上位20位までとか、30位までとかいうやり方で、キー・ワードの選定範囲を指定すると、長い文章についてはキー・ワードが少なめに、短い文章の場合は多めに選ばれてしまうことになる。これを避ける一つの方法としては、累積使用率について、何パーセント以下と指定すれば、文章のボリュームの多少にかかわらず、理論的には、一定のウエイトでキー・ワードが選び出されることになる。

つぎに、さきに掲げた表1でも、あるいは後に掲げる表2・表4などいずれの表においても、その文章の話題の主、小説類についていえば、主人公・ヒロイン・登場人物といったものは、頻度順用語表のきわめて上位を占める傾向が強い。それに対して、話の場面や道具立てに関することばは、話題の主よりは、やや低いところに出てくるという一般的な傾向が認められる。そこで、今回の抄録実験においては、キー・ワードの中で、比較的上位を占めるものを、話題の主になる可能性の高いものという意味で「話題語」と名づけ、それ以外を「場面語」と名づけることにした。そして「話題語」として扱う範囲は、累積使用率が、25%ラインをこえるところまでとした。表1・表2・表4の各表において、見出しに◎印のついたものが、「話題語」と仮定されたキー・ワードであり、○印が「場面語」である。

さて、第三の仮説は、上に述べたように、「話題語」となったキー・ワード

は、小説類ならば主人公やヒロインが含まれるはずの「話題の主」を表わすものであるから、それらが、文章の中で最後に使われているセンテンスは、話の結びになる可能性が強いということである。簡単にいえば、話題語の最終出現文を重視するということである。

以上挙げたキー・ワードについての三つの仮説を整理すると、つぎのようなことになる。

○キー・ワード初出センテンスは、文章の展開の上で重要である。

○キー・ワード、特に話題語は、頻度順用語表の上位を占める。

○話題語の最終出現文は、話の結びになる。

これを前提として、抄録を試みようというのが、今回の抄録実験の大筋である。

かつて、野元菊雄氏は、論文「新聞小説のダイジェスト」^(注5)において、毎月月初めの紙面に出る「前回までのあらすじ」を分析し、「あらすじ」における登場人物の出入りが、原作の流れに対応している点や、話の「真の発端」から「事実の蓄積」が始まり、終末近くになってテンポが早くなる傾向がある点などを指摘している。文章の抄録を行なう以上、このように、話題の提出や切りかえといった原文の流れをとらえるとともに、情報の蓄積の様相も、うまくとらえるものであってほしい。そのためには、やはり、原文におけるキー・ワードの現われ方を、きめ細かく追って行くシステムが必要なのではないかと思う。今回の実験は、そうした方向を旨として設計したものである。

注5) 「言語生活・127号」所収

3. 「藪の中」についての実験例

つぎに掲げるものは、芥川竜之介の「藪の中」の文章についての抄録実験の結果である。この抄録文は、「藪の中」のセンテンス総数315文の中から、その20パーセントに当たる63文が抜き出されている。文頭のアルファベット記号は、そのセンテンスが、「4. 処理過程の概要」に述べるプロセスの、どの段階で抜き出された文であるかを示すものである。

- 1 F 検非違使に問われたる木樵りの物語
- 2 G わたしは今朝いつもの通り、裏山の杉を伐りにまいりました。
- 3 G すると山蔭の藪の中に、あの死骸があったのでございます。
- 4 G 「太刀か何かは見えなかったか？」
- 5 G (が、) 草や竹の落ち葉は、一面に踏み荒されておりましたから、きっとあの男は殺される前に、よほど手痛い働きでもいたしましたのに違いございません。
- 6 F 検非違使に問われたる旅法師の物語
- 7 G あの男は馬に乗った女といっしょに、関山の方へ歩いてまいりました。
- 8 I 男は、——いえ、太刀も帯びておれば、弓矢も携えておりました。
- 9 F 検非違使に問われたる放免の物語
- 10 G これは確か多襄丸という、名高い盗人でございます。
- 11 I この多襄丸というやつは、洛中に徘徊する盗人の中でも、女好きのやつでございます。
- 12 I その月毛に乗っていた女も、こいつがあの男を殺したとなれば、どこへどうしたかわかりません。
- 13 F 検非違使に問われたる媼の物語
- 14 I これは男にも劣らぬくらい、勝気の女でございますが、まだ一度も武弘のほかには男を持ったことはございません。
- 15 F 多襄丸の白状
- 16 I あの男を殺したのはわたしです。
- 17 I しかし女は殺しはしません。
- 18 I わたしはその咄差の間に、たとい男は殺しても、女は奪おうと決心しました。
- 19 I 何、男を殺すなどは、あなた方の思っているように、たいしたことではありません。
- 20 I どうせ女を奪うとなれば、必ず男は殺されるのです。
- 21 I ただわたしは殺す時に、腰の太刀を使うのですが、あなた方は太刀を使わない。
- 22 I しかし男は殺さずとも、女を奪うことが出来れば、別に不足はないわけです。
- 23 I いや、その時の心もちでは、出来るだけ男を殺さずに、女を奪おうと決心したのです。
- 24 I わたしはこれも実をいえば、思う壺にはまったのですから、女一人を残したまま男と藪の中へはいりました。
- 25 I 男はわたしにそう言われると、もう痩せ杉が透いて見える方へ、一生懸命に進んで行きます。
- 26 I 男も太刀を佩いているだけに、力は相当にあったようですが、不意を打たれてはたまりません。

- 27 I わたしは男を片づけてしまうと、今度はまた女のところへ、男が急病を起したらしいから、見に来てくれと言いに行きました。
- 28 I 女は市女笠を脱いだまま、わたしに手をとられながら、蕪の奥へはいつてきました。
- 29 I ところが、そこへ来てみると、男は杉の根に縛られている。
- 30 I わたしはとうとう思い通り、男の命は取らずとも、女を手に入れることは出来たのです。
- 31 I わたしはその上にも、男を殺すつもりはなかったのです。
- 32 I ところが泣き伏した女をあとに、蕪の外へ逃げようとする、女は突然わたしの腕へ、気遣いのようにすがりつきました。
- 33 G しかも切れぎれに叫ぶのを聞けば、「あなたが死ぬか夫が死ぬか、どちらか一人死んでくれ、二人の男に恥を見せるのは、死ぬよりもつらい」と言うのです。
- 34 I わたしはその時猛然と、男を殺したい気になりました。
- 35 G わたしは女と眼を合せた時、たとい神鳴に打ち殺されても、この女を妻にしたいと思いました。
- 36 I 男もそうすればわたしの太刀に、血を塗ることはならなかったのです。
- 37 I が、薄暗い蕪の中に、じっと女の顔を見た殺那、わたしは男を殺さない限り、ここは去るまいと覚悟しました。
- 38 I しかし、男を殺すにしても、卑怯な殺し方はしたくありません。
- 39 I わたしは男の縄を解いた上、「太刀打ちをしろ」と言いました。
- 40 I 男は血相を変えたまま、太い太刀を引き抜きました。
- 41 I わたしは男が倒れると同時に、血に染まった刀を下げたなり、女の方を振り返りました。
- 42 I わたしは、女がどちらへ逃げたか、杉むらの間を探してみました。
- 43 I ことによるとあの女は、わたしが太刀打ちを始めるが早いか、人の助けでも呼ぶために、蕪をくぐって逃げたのかも知れない。
- 44 F 清水寺に来たれる女の懺悔
- 45 I ——その紺の水干を着た男は、わたしを手ごめにしてしまうと、縛られた夫を眺めながら、嘲るように笑いました。
- 46 I (が、)あの盗人に奪われたのでしょ、太刀はもちろん弓矢さえも、蕪の中には見当りません。
- 47 I 夫は、わたしを蔑んだまま、「殺せ」と一言言ったのです。
- 48 I (しかし) 夫を殺したわたしは、盗人の手ごめに遇ったわたしは、いったいどう

すればよいのでしょうか。

49 F 巫女の口を借りたる死霊の物語

50 I — 盗人は妻を手ごめにすると、そこへ腰を下したまま、いろいろ妻を慰め出した。

51 H この男の言うことを真に受けるな。

52 H 「そんな夫に連れ添っているより、自分の妻になる気はないか？」

53 I 盗人にこう言われると、妻はうっとりと顔をもたげた。

54 I (しかし) その美しい妻は、現在縛られたおれを前に、何と盗人に返事をしたか？

55 I (しかし) 妻は夢のように、盗人に手をとられながら、藪の外へ行こうとするとたちまち顔色を失ったなり、杉の根のおれを指さした。

56 I — 妻はそう叫びながら、盗人の腕にすがっている。

57 I 盗人はじっと妻を見たまま、殺すとも殺さぬとも返事をしない。

58 H 「あの女はどうするつもりだ？」

59 H 「殺すか？」

60 I 妻はおれがためらううちに、何か一声叫ぶが早い、たちまち藪の奥へ走り出した。

61 I 盗人は妻が逃げ去った後、太刀や弓矢を取り上げると、一箇所だけおれの縄を切った。

62 H — おれは盗人が藪の外へ、姿を隠してしまう時に、こう呟いたのを覚えている。

63 H おれの前には妻が落した小刀が一つ光っている。

4. 処理過程の概要

今回の抄録実験の処理過程は、大きく3つの段階に分れている。まず、第一の段階は、原文の用語の使用頻度を調査して、その結果からキー・ワードを選定する仕事である。第二の段階は、キー・ワードを目やすにして、原文からセンテンスを抜き出す操作であり、最後の第三段階は、抜き出したセンテンスを、原文の中での出現順に並べたりする編集の作業である。

A) もちろん、その前に、原文からどれだけのセンテンスを抽出して抄録を作るかとか、頻度順位の何位ぐらいまでの頻度調査を行なうかといった、各種の数値などを指定する準備作業が必要である。ここにとりあげた「藪の中」につ

いての抄録実験では、センテンスの抽出は、原文の総センテンス数の20%以内（抽出比）とし、キー・ワードを選定する範囲すなわち「籤の中」の用語の頻度調査の範囲は、累積使用率が40パーセントを越えるところまでとした。実際にコンピュータを使用して処理をする場合には、これらの数値を、あらかじめ、パラメータで指定して、コンピュータの中に入れておくことになる。

4.1. キー・ワードの選定

B) 抄録作成の実質的なプロセスは、文章の用語の使用頻度の調査で始まる。「籤の中」についての調査の結果を示すと、表2の通りである。単語の認定すなわち単位切りをどうするかとか、活用語の変化形をいかに処理するかとかいった問題もあるが、とにかく、このプロセスでは、このようなワード・リストが、メモリーの中に出来あがることになる。

C) このワード・リストの中からキー・ワードを選んでいくわけであるが、表2でわかる通り、このリストの中には、「居る・する・ある・それ・事・とき」といった類の、どんな文章にも使われるような、きわめてありふれたことばが、まぎっている。キー・ワードとして使う以上、対象とする文章に特有なことばでなくては意味がない。この実験についていえば、「籤の中」らしい語を選び出す必要があるわけである。しかし、このような語を、表2の単語の中から、どのようにして選んでいくかは、かなりむずかしい問題ではあるが、少くとも、どんな語彙調査においても、いつも上位を占めるような、きわめてありふれた語は、まず落してしまってさしつかえない。そこで国立国語研究所の「総合雑誌の語彙調査」と「雑誌九十種の語彙調査」における「全体」と「各層別」の語彙表から高頻度語を抜き出してみた。この二つの語彙調査の結果は、いずれも、全体と各層別の使用率順語彙表にまとめられていて、両方の語彙調査を合わせると、十種類の使用率順語彙表^(注6)が得られる。これらの語彙表の中から、使用率1パーミル以上の単語を抜き出し、少くとも五種類の表に共通に出てくる単語のリストを、まず作ってみた。

一方、使用頻度の上では、それほど上位を占めていなくても、いわゆる基本語の類も、特定の文章を特徴づける語ではないので、キー・ワードとしては有

表 2 「葦の中」の頻度順語彙表 (延べ語数2415・異り語数 765)

順位	見出し	度数	累積 度数	累積 使用率 %	順位	見出し	度数	累積 度数	累積 使用率 %	順位	見出し	度数	累積 度数	累積 使用率 %
1	わたし	84	84	3.48	22	太	19	706	29.23	41	何	10	1011	41.86
2	居る	71	155	6.42	23	杉	18	724	29.98	41	申す	10		
3	男	46	201	8.32	24	思ふ	17			41	もの	10		
4	その	37	238	9.86	25	葦	17	758	31.39	41	山	10		
5	する	36	274	11.35	26	唯	16							
6	女	34	308	12.75	26	中	16	790	32.71					
7	あ	33	341	14.12	28	しかし	15	805	33.33					
8	あ	32	373	15.44	29	ない	14	819	33.91					
8	一	32	405	16.77	30	あなた	13							
10	殺す	31	436	18.05	30	来る	13							
11	言ふ	30	466	19.33	30	これ	13							
11	おれ	30	496	20.54	30	人	13	871	36.07					
13	それ	23	519	21.49	34	さ	12							
14	事	22	541	22.81	34	う	12							
14	見	22	563	23.31	34	竹	12	907	37.56					
16	妻	21	584	24.59	37	度	11							
16	時	21	605	25.48	37	馬	11							
16	なる	21	626	26.36	37	ま	11							
16	盗	21	647	26.79	37	見える	11	951	39.38					
20	夫	20	667	28.45	41	も	10							
20	この	20	687	28.45	41	上	10							
						方(カタ)	10							

効でない。この点については、水谷静夫氏は、前掲論文「統計的自動抄録の問題点」において

「一般基本語として使用率に拘らず、キー・ワードとしない意味単位の範囲として『現代雑誌九十種の用語用字（第三分冊）』第一章の『語の基本度の表』に掲げた上位百語を指定する」

という提案をしている。この提案にしたがって、いまの語彙調査結果から得た高頻度語リストに、「語の基本度の表」の上位百語を付け加えると、表3のようになる。（使用目的からいって、「せる・させる・れる・られる」の類は省いてある）。

このプロセスでは、表3にあげたような単語のリストを、テーブル（辞書）として、コンピュータにセットすることになる。これらの単語は、どんな文章にも現われるようなものであって、ある特定の文章や文献の性格とか特徴とかを反映することは、ほとんどない。いわば無性格な語群である。そうしたところから、この種のを「無性格語」と名づけた。

D) キー・ワード選定のための、いちばん主要なステップは、さきに⑧のステップで作成した表2のワード・リストと、いま⑥のステップでセットした無性格語のリスト表3とを照合する作業である。これによって一致したものは、無性格語であるから、キー・ワードとしては、まず失格する。表2において×印のついたものが、それである。

E) しかし、⑨のステップで失格しなかったもの（×印以外）を見渡してみると、キー・ワードとして使っても、あまり効果が期待できない「わたし、おれ、唯、もう」といった代名詞・副詞の類が残っている。一般に、こうした処理の場合、キー・ワードとしては、文の中核的な意味をになうものでなくては、有効性がない。これについて水谷氏は、前掲「抄録を作る機械」^(注2)において、キー・ワードの選定範囲を「名詞・形容動詞語幹類・動詞・形容詞」に限ることを提唱している。しかし、名詞の中でも、たとえば「上・下・右・左・前・うしろ・あと・さき・間」というような関係概念を表わす類などは落した方がよい。そこで、この種の名詞や代名詞・副詞・連体詞・接続詞・感動詞の類を削除していくことになる。表2で△印のついているものが、それである。

表 3 無性格語の表

あまり	事	強い	方 (ホウ)
あ る	この	的 (テキ)	僕
い う	これ	できる	程
一	三	出る	前
今	さん	度	また
居 る	三十	どう	万
う ち	氏 (シ)	時 (トキ)	見る
円	四 (シ)	ところ	目
お	しかし	とも	持つ
多 い	七	とる	もの
大きい(な)	自分	無い	問題
置 く	しまう	中 (ナカ)	やる
於 く	者 (シヤ)	何	行く
同 じ	知る	なる	よい
思 う	十	二	よう
居 (オ) る	する	日 (ニチ)	よる
会 (カイ)	生活	二十	四 (ヨン)
方 (カタ)	千	日本	四十
月 (ガツ)	そう	人 (ニン)	等 (ラ)
彼	そして	年 (ネン)	零
考える	その	はいる	六
聞 く	それ	場合	分かる
九	第	八	訳 (ワケ)
くらい	対する	日 (ヒ)	わたくし
来 る	出す	人 (ヒト)	
五	達	ひとつ	
こう	為	ひとり	
五 十	つく	百	

しかし、この処理を全く機械的に行なうとなると、この種の語のリストを作って照合するか、入力前のデータの各語に品詞づけなどの作業をしておくか、あるいは、品詞等の自動認定プログラムを開発しなくてはならない。どの方法も、かなり手間を食ううえ、実際にこの段階で削除すべき語の数というものは、多くても10語程度のことであるから、ここには、やはり人間が介入した方が手っとり早い。すなわち表2において×がついている見出し以外のものを、一度、アウト・プットして、△印のついているような語を人間が削除した上で再入力するわけである。

以上の処理によって、表2の×印△印のついた語が、キー・ワード候補から脱落し、頻度順に挙げると「男・女・殺す・妻・盗人・夫・太刀・杉・藪・竹・馬・見える・山」の13語が残ってくる。これが、この実験で使用するキー・ワードである。

なお、今回の実験では、さきに「3・キー・ワードについての仮説」において述べたように、キー・ワードのうち、累積使用率が25パーセントのラインを越えるところまでのものを「話題語」と指定し、それ以外を「場面語」としたので、「藪の中」においては、「男・女・殺す・妻・盗人」の五語が「話題語」となり、それ以外は「場面語」ということになる。

注6) 国立国語研究所報告21「現代雑誌九十種の用語用字(第一分冊)」の「全体／評論・芸文／庶民／実用・通俗科学／生活・婦人／娯楽・趣味」の6表と、同研究所報告12「総合雑誌の用語(前編)」の「全体／一層／二層／三層」の4表、計10表。

4.2 センテンスの抽出

F) 「藪の中」の文章は、全体が7つのパラグラフに分かれ、その一つ一つには、タイトルがついている。一般に、章節のタイトルというものは、話題の重要な変化や、場面の大きな展開を示していることが多い。したがって、章節にタイトルがついている文章の場合には、抄録文の中に、タイトルを採用しておくことは、多くの場合、きわめて有利であると考えられる。

こうした見地から、この実験では、章節のタイトルとなっているセンテンス

は、無条件で採用することにした。無論、タイトルをもたないもの場合は、このプロセスはスキップする。

G) すでに「3・キー・ワードについての仮説」において述べた通り、今回の実験では、キー・ワード初出センテンスは、優先的に抄出することになっている。したがって、さきあげた13個のキー・ワードが最初に出現するセンテンスを、まず抜き出していく。その結果、抽出されてくる文は、「籤の中」において、話題の提出や転換あるいは場面の変換をになっていると予想されるセンテンスである。

H) つぎは、「話題語」の最終出現センテンスの抽出である。「籤の中」での実験結果では、「51・58・59・62・63」の5文が、それに当る。これらは、各々の話題語についての話の結びと想定されるものである。

I) センテンス抽出の、つぎの作業は、「話題語を少なくとも1個含み、かつ2個以上のキー・ワードを含む文」を抜き出すことである。すでに「3・キー・ワードについての仮説」において述べた通り、「話題語」が話題の主ないしは話の核を表わすものと仮定され、「場面語」が話の背景や道具立てに関する語と想定されている以上、このようなセンテンスは、話の筋の展開や場面の移り変わりの中で、かなり重要なセンテンスではないかと想像されるものである。

J) 「籤の中」の場合には、上の①の段階までに抄出したセンテンスの数が、72文になってしまい、当初予定した原文の20%抽出という枠を大きくオーバーしてしまう。その場合には、頻度順位の最下位のキー・ワードから順に削除してキー・ワードの個数を減らしつつ、⑥以降の操作を繰り返して、抄出文の数を減じていく。「籤の中」の場合には、下の方から「山・見える・馬・竹」の4語をキー・ワードからはずしたとき、抄出文の割合が、当初予定した20パーセント63文となった。したがって、この実験で最終的に使用したキー・ワードは表2にゴチック活字で示した「男・女・殺す・妻・盗人・夫・太刀・杉・籤」の9語、頻度順位25位、累積使用率31.39%までのものということになる。

K) システムとしては、①までの操作の結果、今の例とは逆に、抄出文数が予定した抽出比に達しない場合が生じてくる。この場合には、キー・ワードの出現回数が多いセンテンスから順に抽出していく。

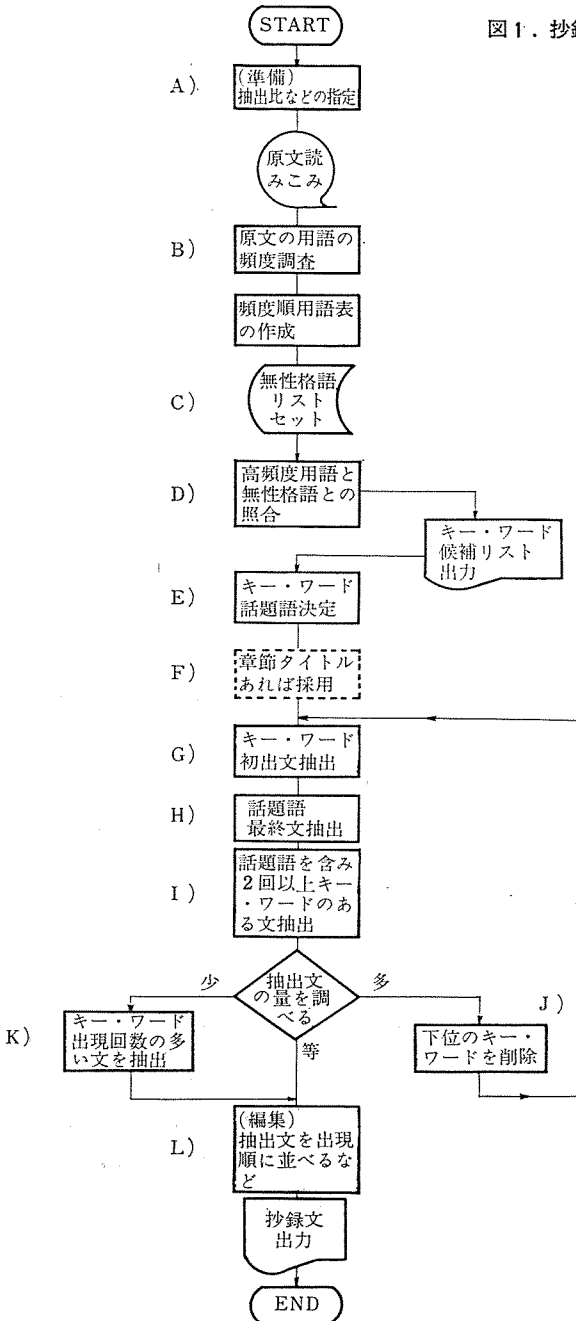
4.3. 編集作業

L) 最後は編集の仕事であるが、これについては、今回の実験では、抄出したセンテンスを出現順に並べる作業と、対話部分から抽出されてきたセンテンスに「 」をつける処理をした程度である（ただし、「籤の中」の場合には、本人の語り部分ではない「他人の話を用いた部分」に「 」をつけてある）。

しかし、先行文を抄出されていないセンテンスの場合、文頭の接続詞などは「接続詞リスト」とでも照合して削除しておいた方が、出来あがりがかきれいになる。前掲実験例において、パーレンで囲んで示したものが、それに当たる。また、同様なことであるが、たとえば実験例の62番のセンテンスのように、先行文のない場合の指示語の処理なども問題になろう。「籤の中」では、例がなかったが、「AはBに話しかけた。『いつ御出発ですか』』」といったような個所で、先行の「AはBに話しかけた」の部分だけが抽出されてしまうと、抄録文の中で浮き上がってしまう。こうした場合には、やはり『いつ御出発ですか』までを含めて一文として扱うような処置が必要であろう。いずれにしても、編集の手法については、もっとキメ細かく考えていかななくてはならない。

以上述べた抄録実験の処理過程、思考過程を流れ図の形で示すと、図1のようになる。この抄録方式のかわっている点は、さきに第2節および第3節で述べたように、キー・ワード密度による方式を採らず、キー・ワードの現われ方を重視したところにある。しかし、章節タイトルや、キー・ワード初出文あるいは話題語の最終出現文を重点的ないしは優先的に処理したあとの①①①などのあたりは、キー・ワード密度による方式で進めるのも一法かもしれない。文章の筋道や論理の一貫性が、一応保証されているような論文とか評論の類の文章、あるいは、さきに第一節で述べたようにキー・ワード密度が大きく乱れる対話部分を、持たない文章などの場合には、今回実験したキー・ワードの現われ方によって進める抄録方式と、きわめてドライな密度方式とを併用するのも一つの考え方ではないかと思うのである。

図 1. 抄録処理の流れ図



5. 「高瀬舟」についての実験

以下に掲げるのは森鷗外の「高瀬舟」についての実験例である。この抄録には、「高瀬舟」の総センテンス数 213 文の中から、その約20パーセントに当たる42文が抽出されている。この場合のキー・ワードは、表4に示す「喜助・弟・庄兵衛・罪人・島・顔・抜く・同心・手・為事・高瀬舟」の11語であり、「喜助・弟・庄兵衛」が「話題語」、その他が「場面語」ということになる。また、この「高瀬舟」の処理においては、さきの「籟の中」とは逆に、①のステップまでの抄出文数が20%を下まわったため、図1の「流れ図」の⑩のステップにジャンプして行き、キー・ワード出現回数の最も多い文として、4回出ている「14・35」の文が採用されている。なお、この作業のデータには、国立国語研究所言語計量調査室の作成した「KWIC索引」^(注7)を使用した。

注7) 石綿敏雄「KWICの設計」計量国語学60号

- 1 G 高瀬舟は、京都の高瀬川を上下する小舟である。
- 2 G 徳川時代に京都の罪人が遠島を申し渡されると、本人の親類が牢屋敷へ呼び出されて、そこで暇乞をすることが許された。
- 3 G それを護送するのは、京都町奉行の配下にゐる同心で、此同心は罪人の親類の中で、主立った一人を大阪まで同船させることを許す慣例であった。
- 4 G それは名を喜助と云って、三十歳ばかりになる住所不定の男である。
- 5 G 護送を命ぜられて、一しょに舟に乗り込んだ同心羽田庄兵衛は、只喜助が弟殺しの罪人だと云ふことだけを聞いてゐた。
- 6 I 夜舟で寝ることは、罪人にも許されてゐるのに、喜助は横にならうとはせず雲の濃淡に従って、光の増したり減じたりする月を仰いで黙ってゐる。
- 7 I 庄兵衛はまともには見てゐぬが、終始喜助の顔から目を離さずゐる。
- 8 G それは喜助の顔が縦から見ても、横から見ても、いかにも楽しさうで、若し役人に対する気兼ねがなかったら、口笛を吹きはじめるとか、鼻歌を歌ひ出すとかしきうに思はれたからである。
- 9 G 罪は弟を殺したのださうだが、よしや其弟が悪い奴で、それをどんな行掛りになつて殺したにせよ、人の情として好い心持はせぬ筈である。
- 10 I 庄兵衛がためには喜助の態度が考へれば考へる程わからなくなる。
- 11 I 「はい」と云ってあたりを見廻した喜助は、何事をお役人に見咎められたので

はないかと氣遣ふらしく、居ずまひを直して庄兵衛の氣色を伺った。

12G 「実 はな、己は先刻からお前の島へ往く心持ちが聞いて見たかったのだ。」

13G 「(それに) わたくしはこんなにかよわい体ではございますが、つひぞ病氣をいたしたことはございませんから、島へ往ってから、どんなつらい為事をして、体を痛めるやうなことはあるまいと存じます。」

14K 「島へ往って見ますまでは、どんな為事が出来るかわかりませんが、わたくしは此二百文を島でする為事の本手にしようと楽しんでをります。」

15G こう云ひ掛けて、喜助は胸に手を当てた。

16 I 庄兵衛は彼此初老に手の届く年になってゐて、もう女房に子供を四人生ませてゐる。

17 I 庄兵衛は五節句だと云つては、里方から物を貰ひ、子供の七五三の祝だと云つては、里方から子供に衣類を貰ふのでさへ、心苦しく思つてゐるのだから暮しの穴を填めて貰つたのに氣が付いては好い顔はしない。

18 I 庄兵衛は今喜助の話を受けて、喜助の身の上をわが身の上引き比べて見た。

19 I 喜助は為事をして給料を取つても、右から左へ人手に渡して亡つてしまふと云つた。

20 I 喜助は世間で為事を見附けるのに苦しんだ。

21 I それを今日の前で踏み止まってくれるのが此喜助だと、庄兵衛は氣が付いた。

22 I 庄兵衛は今さらのやうに驚異の目を睜つて喜助を見た。

23 I 此時庄兵衛は空を仰いでゐる喜助の頭から毫光がさすやうに思った。

24 I 庄兵衛は喜助の顔をまもりつつ又、「喜助さん」と呼び掛けた。

25 I 「はい」と答えた喜助も「さん」と呼ばれたのを不審に思ふらしく、おそるおそる庄兵衛の氣色を覗いた。

26 I 「(すると) 弟は真蒼な顔の、両方の頬から腮へ掛けて血に染つたのを挙げて、わたくしを見ましたが、物を言ふことが出来ませぬ。」

27 I 「わたくしにはどうも様子がわかりませんので、『どうしたのだい、血を吐いたのかい』と云つて、傍へ寄らうといたすと、弟は右の手を床に衝いて、少し体を起しました。」

28G 「これを旨く抜いてくれたら己は死ぬるだらうと思つてゐる。」

29 I 「弟が左の手を弛めるとそこから又息が漏ります。」

30 I 「わたくしはなんと云はうにも、声が出ませんので、黙つて弟の喉の創を覗いて見ますと、なんでも右の手に剃刀を持って、横に笛を切つたが、それでは死に切

れなかったので、其儘剃刀を、削るやうに深く突っ込んだものと見えます。

31 I 「わたくしはそれだけの事を見て、どうしようと云ふ思案も附かずに、弟の顔を見ました。」

32 I 「弟は怨めしさうな目附をいたしましたが又左の手で喉をしっかり押へて、『医者がなんになる、あゝ苦しい、早く抜いてくれ、頼む』と云ふのでございます」

33 I 「わたくしは途方に暮れたやうな心持になって、只弟の顔ばかり見てをります」

34 I 「弟は衝いてゐた右の手を放して、今まで喉を押へてゐた手の肘を床に衝いて、横になりました。」

35 K 「わたくしは剃刀を抜く時、手早く抜かう、真正に抜かうと云ふだけの用心はいたしました、どうも抜いた時の手応は、今まで切れてゐなかつた所を切つたやうに思はれました。」

36 I 「婆あさんが行つてしまつてから、気が付いて弟を見ますと、弟はもう息が切れてをりました。」

37 I 「それから年寄衆がお出になつて、役場へ連れて行かれますまで、わたくしは剃刀を傍に置いて、目を半分あいた儘死んでゐる弟の顔を見詰めてゐたのでございます。」

38 I 少し俯向き加減になつて庄兵衛の顔を下から見上げて話してゐた喜助は、かう云つてしまつて視線を膝の上に落した。

39 I 弟は剃刀を抜いてくれたら死なれるだらうから、抜いてくれと云つた。

40 H しかし其儘にして置いても、どうせ死ななくてはならなかつた弟であつたらしい。

41 H 喜助は其苦を見てゐるのに忍びなかつた。

42 H さうは思つても、庄兵衛はまだどこやらに腑に落ちぬものが残つてゐるので、なんだかお奉行様に聞いて見たくてならなかつた。

6. キー・ワードを含む文の分布

自動抄録を試みる場合、どんな手法で行なうにしても、結局は、キー・ワードを手掛りにして原文から一定量のセンテンスを抜き出していくという基本的手順が変らない以上、キー・ワードの個数や、その使用頻度が、抄出センテンスの量に、どのように影響するかということは、きわめて重要な問題である。

今回試みたような、キー・ワードの現われ方を手掛りにして進めていく場合は、もちろんであるが、文におけるキー・ワードの密度の大小によつて文の抽

表 4 キー・ワードの表

高瀬舟 (森鷗外) 延べ語数 2484 / 異り語数 846 / 総文数 213											
順位	度数	使用率	累積 度数	累積 使用率	見出し	抽出 文数	抽出比	新出 文数	累積 文数	累積 抽出比	
		%		%			%			%	
10	33	1.33	494	19.86	喜 助 ◎	31	14.55		31	14.55	
13	27	1.09			弟 ◎	25	11.74	+25	56	26.29	
13	27	1.09	608	24.48	庄兵衛 ◎	27	12.68	+15	71	33.33	
23	14	0.56	811	32.65	罪 人 ○	14	6.57	+12	83	38.97	
26	13	0.52	837	33.70	鳥 ○	12	5.63	+12	95	44.60	
33	11	0.44			顔 ○	11	5.16	+ 2	97	45.54	
33	11	0.44			手 ○	10	4.69	+ 6	103	48.36	
33	11	0.44	930	37.44	抜 く ○	8	3.76	+ 2	105	49.30	
36	10	0.40	970	39.05	同 心 ○	6	2.82	+ 2	107	50.23	
42	8	0.32			為 事 ○	6	2.82	+ 2	109	51.17	
42	8	0.32	1068	43.00	高瀬舟 ○	8	3.76	+ 3	112	52.58	
葦の中 (芥川竜之介) 延べ語数 2415 / 異り語数 765 / 総文数 315											
3	46	1.90	201	8.32	男 ◎	44	13.97		44	13.97	
6	34	1.41	308	12.75	女 ◎	31	9.84	+19	63	20.00	
10	30	1.28	436	18.05	殺 す ◎	29	9.21	+14	77	24.44	
16	21	0.87			妻 ◎	20	6.35	+18	95	30.16	
16	21	0.87	647	26.79	盗 人 ◎	21	6.67	+12	107	33.97	
20	20	0.83	687	28.45	夫 ○	20	6.35	+14	121	38.41	
22	19	0.79	706	29.23	太 刀 ○	19	6.03	+ 8	129	40.95	
23	18	0.75	724	29.98	杉 ○	18	5.71	+13	142	45.08	
25	17	0.70	758	31.39	藪 ○	17	5.40	+ 8	150	47.62	
34	12	0.50	907	37.56	竹 ○	12	3.81	+ 3	153	48.57	
37	11	0.46			馬 ○	10	3.17	+ 6	156	49.52	
37	11	0.46	951	39.38	見える ○	9	2.86	+ 3	159	50.48	
41	10	0.41	1011	41.86	山 ○	10	3.17	+ 3	162	51.43	
城の崎にて (志賀直哉) 延べ語数 1681 / 異り語数 696 / 総文数197											
12	19	1.13	450	26.77	いもり ○	17	8.54		17	8.54	
15	14	0.83	497	29.57	蜂 ○	13	6.53	+13	30	15.08	
16	13	0.77	536	31.89	ねずみ ○	13	6.53	+13	43	21.61	
19	12	0.71	584	34.74	石 ○	12	6.03	+11	54	27.14	
23	11	0.65	628	37.36	死 ぬ ○	11	5.53	+ 6	60	30.15	
29	8	0.48	694	41.28	書 く ○	7	3.52	+ 6	66	33.17	
たき火 (志賀直哉) 延べ語数 2398 / 異り語数 802 / 総文数 278											
2	77	3.21	186	7.76	K ◎	74	26.62		74	26.62	

順位	度数	使用率	累積 度数	累積 使用率	見出し	抽出 文数	抽出比	新出 文数	累積 文数	累積 抽出比	
		%		%			%			%	
9	25	1.04	504	21.02	S	◎	24	8.63	+17	91	32.73
13	21	0.88	572	23.85	山	◎	16	5.76	+8	99	35.61
14	20	0.83	612	25.52	妻	◎	20	7.19	+15	114	41.01
23	12	0.50	771	32.15	見える	○	12	4.32	+6	120	43.17
26	11	0.46	804	33.53	登る	○	10	3.60	+3	123	44.24
29	10	0.42	864	36.03	雪	○	10	3.60	+7	130	46.46
35	9	0.38			小屋	○	8	2.88	+2	132	47.48
35	9	0.38			話	○	8	2.88	+3	135	48.56
35	9	0.38	936	39.03	舟	○	9	3.24	+6	141	50.72
43	8	0.33			気持	○	8	2.88	+5	146	52.52
43	8	0.33			静か	○	8	2.88	+2	148	53.24
43	8	0.33			たき火	○	8	2.88	+5	153	55.04
43	8	0.33			不思議	○	8	2.88	+2	155	55.76
43	8	0.33			水	○	7	2.52	+2	157	56.47
43	8	0.33	1024	42.70	森	○	8	2.88	+1	158	56.83

蜘蛛の糸（芥川竜之介）延べ語数 826／異り語数 307／総文数52

3	17	2.06	59	7.14	韃陀多	◎	17	32.19		17	32.19
5	14	1.69			糸	◎	14	26.92	+9	26	50.00
5	14	1.69	116	14.04	蜘蛛	◎	14	26.92	+1	27	51.92
8	13	1.57	142	17.19	地獄	◎	12	23.08	+6	33	63.46
11	10	1.21			極楽	◎	10	19.28	+5	38	73.08
11	10	1.21	204	24.70	登る	◎	9	17.31	+1	39	75.00
16	9	1.09	231	27.97	池	◎	9	17.31	+3	42	80.77
19	8	0.97	247	29.90	血	○	7	13.46	0	42	80.77
21	7	0.85	261	31.60	釈迦	○	7	13.46	+1	43	82.69
23	6	0.73	279	33.78	罪人	○	6	11.54	+1	44	84.62
26	5	0.61			落ちる	○	5	9.62	+1	45	86.54
26	5	0.61			針	○	4	7.69	+2	47	90.38
26	5	0.61	309	37.41	光る	○	5	9.62	0	47	90.38
32	4	0.48			男	○	4	7.69	0	47	90.38
32	4	0.48			きれる	○	4	7.69	0	47	90.38
32	4	0.48	361	40.31	蓮	○	4	7.69	0	47	90.38

出を進めていく方式の場合も、最終的には、抄録文の量が制限されている以上キー・ワード密度の採用基準点をどこに置くか、もっと具体的にいえば、キー・ワード密度何パーセント以上のセンテンスを採用するかを決定するさいにはやはり、キー・ワードが、どれだけセンテンスに、どのように分布している

かの見通しがなくては、合理的な基準点がわり出せないはずである。

そこで、いくつかの文章について、今まで述べてきた抄録実験において採用した方法で、累積使用率が40パーセント・ラインを超えるところまでの範囲から、キー・ワードを選定し、それが含まれるセンテンスの数を示すと、表4のようになる。この表の見出しの左側の方は、頻度順用語表の中から、見出しに掲げた語に関する数値のみを抜き出して示したものである。したがって、順位や累積度数は、この表においては連続しない。

表4の見出し語の右側の数値は、これらの語をキー・ワードとして、これを含むセンテンスを、もし抽出した場合、どれだけのセンテンスが原文から抽出されてくるかを示したものである。このうち、「抽出比」というのは、「各々のキー・ワードすなわち見出語を含む文」の、総センテンス数に対する比率であり、「新出文数」というのは、当該キー・ワードを追加することによって、新たに加わるセンテンスの数である。また「累積文数」というのは、「新出文数」を累計したもので、最上位のキー・ワード（見出し）から、当該キー・ワードまでで抽出されうるセンテンスの数であり、「累積抽出比」は、「累積文数」の全センテンスに対する比率である。したがって、「累積抽出比」は、最上位のキー・ワードから、当該キー・ワードまでで、全体の何パーセントの文が抽出されうるかを示すものである。

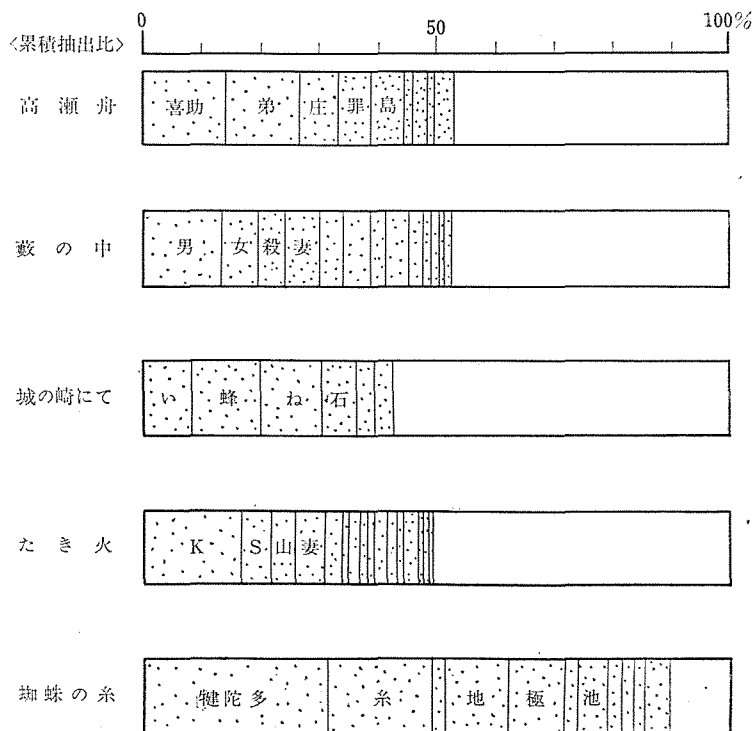
表4からわかる通り、キー・ワードの使用頻度順位は、抽出文数の順位と、ほぼ、パラレルになっている。したがって、使用頻度の高いキー・ワードは、やはり、多くの文に含まれていて、多くの文を抽出してくるのに対して、使用頻度の低いキー・ワードは、抽出してくる文の数も少ないということを物語っている。この事実は、きわめて常識的な結果ではあるが、キー・ワードを、語の使用頻度を手掛りにして選定していく手法の合理性を裏づけるものとして重要である。

本来、キー・ワードというものは、理想的には、話題の主あるいは話の重要な場面や道具立てに関するものが、すべて選ばれているはずだから、それを含むセンテンスも、当然、話の核となりうるものであり、話の展開の軸であるはずである。こうした観点からいえば、常識的には、キー・ワードは、少いより

は多い方が話題を広くおおい、抄録の中に、話の多彩な展開を描き出しうる点で有利だと考える考え方もあり得よう。

そこで、表4の「新出文の累積比率」の増加のしかたを、グラフにしてみると、図2のようになる。このグラフは、頻度順上位のキー・ワード群が、きわめて高い割合で、センテンスを抽出するのに対して、下位のキー・ワードは、新出文の抄出には、ほとんど役立っていないことを、はっきりと示している。

図2. キー・ワードのセンテンス抽出能力



したがって、なるべく多くの文を選び出すということだけを目ざすならば、頻度の高い語群の中からキー・ワードを得るように心がけた方が、キー・ワードの数を増すよりは効果的だということになる。このことは、一方からみれば、頻度順位の下のキー・ワードは、たとえ、ふやしても、それが新たに抽出するセンテンス（新出文）の数は、ごくわずかしかないということでもある。

さらに言えば、抄録のボリュームを強く制限する場合には、キー・ワードの

数をしぼるよりは、頻度順位の上位の語群からのキー・ワードの選定を厳しく制限した方が効果的であるという結論も出てくる。逆にキー・ワードのおおう話題の範囲を広げるために、頻度順の下位の方にまでキー・ワード選定範囲を広げてみても効果がないということを裏づけるものであり、「キー・ワードが多い方が話題を広くおおうのではないか」という予想は必ずしも当たらないようである。

また、図2のグラフでもわかる通り、単純に、キー・ワード含有文を抄出していったのでは、キー・ワードの数を、かなり限定したとしても、とても「抄録」ということにはならない。極端な例では、「蜘蛛の糸」の場合など、上位2語を採っただけで、すでに全文の5割のセンテンスが抽出されてしまう。今回の実験のように累積使用率40パーセントあたりまでを、キー・ワードの選定範囲にすると、単純に含有文を抽出していった場合には、少ないものでも3割以上の文が抜き出され、多い場合には9割ものセンテンスが抽出されてくる。これを避けるために採用されるのが、キー・ワード密度による方式とか、初出センテンスや最終出現文の重視とかいった手だてであろうが、すくなくとも、こうした手法の有効性を、はっきりさせるためには、ここに採りあげたようなデータを、もっと広く集めて、分析してみる必要があるのではないかと考える。

7. 高頻度語を含む文

前節において、高頻度語の中から無性格語を削除する方法によって選定したキー・ワードは、それ自身の出現頻度と、それらが抽出する文の数との間に、ほぼ一定の関係すなわちキー・ワードの頻度数が少なくなるにしたがって抽出文数も減ってくるという関係をもっている点を指摘した。これは、言いかえれば、キー・ワードの頻度順位と、抽出文の度数順位とは密接な関係にあるということであり、この点からみれば、P・H・ルーンの提唱したキー・ワードの選定方式にも、一応の合理性が認められることを述べた。

しかし、自動抄録におけるキー・ワードの役割りは、あくまでも原文において中核的な働きをしているセンテンスを抽出する点にあり、用語の頻度にもと

づいてキー・ワードを選ぶということは、いわば便宜的な手段に過ぎない。結果的には、この方法でキー・ワードを選定しても、そう問題はないとしても、この手法で選ばれたキー・ワード群が、確実に原文から主要センテンスを抽出しようという理論的な根拠は、なお薄弱であるといわなければならない。

そこで、P・H・ルーンの方法をはじめとして、従来の自動抄録において、キー・ワードを選定する基盤となっている高頻度語群というものが、センテンスの抽出において、どのような動きをするものか検討してみたいと思う。

表5・表6・表7は、それぞれ「城の崎にて」「高瀬舟」「たき火」の文章における高頻度語群と、それらを含む文との関係を示したものである。各表の見出し語の左側の数値は、各々の文章の用語の頻度順語彙表の形をなすものである。右側の数値は、それぞれの見出し語が含まれるセンテンスについての数値であり、自動抄録処理に則していえば、各見出し語が原文から抽出してくる文の数についての情報である。各表で、「文数」というのは、見出し語が含まれている文の数（異なり数）すなわち抽出しようる文の数を示したものである。

「抽出比」は「文数」の「総文数」に対する百分比を表わす。したがって、「抽出比」は、その見出し語によって、もし単純にセンテンスの抽出を行なった場合、総文数の何パーセントが抽出しようるかを示すことになる。各表で「新出文数」というのは、その見出し語によって新たに抽出されるセンテンスの数のことであり、「累積文数」は「新出文数」を累計したものである。したがって「累積文数」は、頻度順最上位の見出し語から、当該見出しまでのすべての語が抽出しようる文の数を示すことになる。また「累積抽出比」は「累積文数」の「総文数」に対する百分比であり、これは最上位の見出しから当該見出しまで、全体の何パーセントの文が抽出されようるかを示している。

細かい計算をしてみるまでもなく、これらの表から、いずれの文章の場合も、用語の「使用率」が減少するにしたがって、抽出文の「抽出比」も段階的に減少する一方、用語の「累積使用率」が伸びるにしたがって、抽出文の「累積抽出比」も伸びていくという密接な関係を読みとることができる。この点が、語の使用頻度を手がかりにしてキー・ワードを選定し、これによって文の抽出を進めていく抄録法の、妥当性を裏づける一つの論拠であろうが、しかし、こ

表 5 高瀬舟（森鷗外）における高頻度語と抽出文の関係

<延べ語数2484・異り語数846・総文数213>

用 語					見 出 し	抽 出 文				
順位	度数	使用率	累積 度数	累 積 使用率		抽出 文数	抽出比	新出 文数	累積 文数	累 積 抽出比
1	73	2.94%	73	2.94%	する	59	27.70%		59	27.70%
2	72	2.90	145	5.84	ゐる	62	29.11	+39	98	46.01
3	59	2.38	204	8.21	ある	54	25.35	+31	129	60.51
4	57	2.29	261	10.51	いふ	50	23.47	+21	150	70.42
5	54	2.17	315	12.68	事	46	21.60	+8	158	74.18
6	42	1.69	357	14.37	見る	39	18.31	+8	164	77.00
7	36	1.45	393	15.82	わたくし	34	15.96	+6	170	79.81
8	34	1.37			思ふ	32	15.02	+8	178	83.57
8	34	1.37	461	18.56	なる	31	14.55	+5	183	85.92
10	33	1.33	494	19.86	喜助	◎ 31	14.55	+6	189	88.73
11	30	1.21			それ	29	13.62	+2	191	89.67
11	30	1.21	554	22.30	ない	29	13.62	+4	195	91.55
13	27	1.09			弟	◎ 26	12.21	+1	196	92.02
13	27	1.09	608	24.48	庄兵衛	◎ 27	12.68	0		
15	26	1.05	634	25.12	ござる	25	11.75	+4	200	93.90
16	25	1.01	659	26.53	その	23	10.80	+3	203	95.31
17	24	0.97	683	27.50	もの	21	9.86	0		
18	20	0.81	703	28.30	この	20	9.39	+2	205	96.24
19	19	0.76	722	29.07	行く	17	7.98	+1	206	96.71
20	17	0.68	739	29.75	人	15	7.04	0		
21	15	0.60			開く	13	6.10	0		
21	15	0.60	769	30.96	できる	15	7.04	0		
23	14	0.56			いたす	13	6.10	0		
23	14	0.56			これ	14	6.57	0		
23	14	0.56	811	32.65	罪人	○ 14	6.57	0		
26	13	0.52			島	○ 13	6.10	+3	209	98.12
26	13	0.52	837	33.70	目	12	5.63	0		
28	12	0.48			来る	12	5.63	+1	210	98.59
28	12	0.48			自分	12	5.63	0		
28	12	0.48			つく	11	5.16	0		
28	12	0.48			どう	10	4.68	0		
28	12	0.48	897	36.11	時	11	5.16	0		
33	11	0.44			顔	○ 10	4.69	0		
33	11	0.44			手	○ 9	4.23	0		
33	11	0.44	930	37.44	抜く	○ 7	3.29	0		
36	10	0.40			くれる	9	4.23	0		
36	10	0.40			そう	9	4.23	0		

用 語					見 出 し	抽 出 文				
順位	度数	使用率	累 積 度数	累 積 使用率		抽出 文数	抽出比	新出 文数	累 積 文数	累 積 抽出比
36	10	0.40			出す	10	4.69	0		
36	10	0.40	970	39.05	同心	6	2.82	0		
40	9	0.36			しかし	9	4.23	0		
40	9	0.36	988	39.77	やる	9	4.23	0		
42	8	0.32			うち	8	3.76	0		
42	8	0.32			かう	8	3.76	0		
42	8	0.32			考へる	7	3.29	0		
42	8	0.32			為事	6	2.82	0		
42	8	0.32			しまふ	8	3.76	0		
42	8	0.32			高瀬舟	8	3.76	0		
42	8	0.32			申す	8	3.76	0		
42	8	0.32			持つ	8	3.76	0		
42	8	0.32			をる	8	3.76	+1	211	99.06
42	8	0.32			ところ	7	3.29	0		
52	7	0.28			剃刀	6	2.82	0		
52	7	0.28			死ぬ	7	3.29	0		
52	7	0.28			そこ	7	3.29	0		
52	7	0.28			只	7	3.29	0		
52	7	0.28			出る	7	3.29	0		
52	7	0.28			二百	7	3.29	0		
52	7	0.28			又	7	3.29	0		
52	7	0.28	1124	45.25	文(モン)	7	3.29	0	211	99.06

表 6 城の崎にて(志賀直哉)における高頻度語と抽出文の関係

<延べ語数 1681/異り語数 696/総文数197>

用 語					見 出 し	抽 出 文				
順位	度数	使用率	積 累 度数	積 累 使用率		抽出 文数	抽出比	新出 文数	積 累 文数	積 累 抽出比
1	68	4.05	68	4.05	自分	61	30.96		61	30.96
2	67	3.99	135	8.03	する	55	27.92	+32	93	47.21
3	61	3.63	196	11.66	ゐる	54	27.41	+25	118	59.90
4	46	2.74	242	14.40	それ	41	20.81	+12	130	65.99
5	39	2.32	281	16.72	事	32	16.24	+3	133	67.51
6	30	1.78	311	18.50	なる	26	13.20	+8	141	71.57
7	28	1.67	339	20.17	思ふ	26	13.20	+8	149	75.63
8	26	1.55	365	21.71	有る	24	12.18	+7	156	79.19
9	25	1.49	390	23.20	ない	25	12.69	+5	161	81.73
10	21	1.25	411	24.45	言ふ	20	10.15	+1	162	82.23

用 語					見 出 し	抽 出 文				
順位	度数	使用率	累積 度数	累積 使用率		抽出 文数	抽出比	新出 文数	累積 文数	累積 抽出比
11	20	1.19	431	25.64	その	20	10.15	+1	163	82.74
12	19	1.13	450	26.77	いもり ○	16	8.12	+6	169	85.79
13	18	1.07	468	27.84	良い	17	8.62	+3	172	87.31
14	15	0.89	483	28.73	見る	14	7.11	0		
15	14	0.83	497	29.57	蜂 ○	13	6.60	+2	174	88.32
16	13	0.77			しかし	13	6.60	+1	175	88.83
16	13	0.77			ねずみ ○	13	6.60	+1	176	89.34
16	13	0.77	536	31.89	もの	13	6.60	0		
19	12	0.71			石 ○	12	6.09	+3	179	90.86
19	12	0.71			来る	12	6.09	0		
19	12	0.71			行く	11	5.58	0		
19	12	0.71	584	34.74	前	12	6.09	+1	180	91.37
23	11	0.65			気	10	5.08	0		
23	11	0.65			死ぬ ○	11	5.58	0		
23	11	0.65			どう	11	5.58	0		
23	11	0.65	628	37.36	ところ	10	5.08	0		
27	9	0.54			さう	9	4.57	0		
27	9	0.54	646	38.43	もう	9	4.57	+2	182	92.39
29	8	0.48			書く ○	7	3.55	0		
29	8	0.48			三	7	3.55	0		
29	8	0.48			しまふ	8	4.06	0		
29	8	0.48			ヒラヒラ	2	1.02	0		
29	8	0.48			まったく	8	4.06	0		
29	8	0.48	694	41.28	まま	7	3.55	0		
35	7	0.42			いか	7	3.55	+1	183	92.89
35	7	0.42			忙しい	6	3.05	0		
35	7	0.42			いま	7	3.55	0		
35	7	0.42			動く	6	3.05	0		
35	7	0.42			考へる	7	3.55	0		
35	7	0.42			気持	6	3.05	0		
35	7	0.42			殺す	6	3.05	0		
34	7	0.42			さびしい	7	3.55	+2	185	93.91
35	7	0.42			下	7	3.55	0		
35	7	0.42			静か	7	3.55	0		
35	7	0.42			そんな	7	3.55	0		
35	7	0.42			付く	7	3.55	0		
35	7	0.42			出る	6	3.05	+1	186	94.42
35	7	0.42			流れ	7	3.55	0		
35	7	0.42			一つ	6	3.05	0		

用 語					見 出 し	抽 出 文				
順位	度数	使用率	累積 度数	累 積 使用率		抽出 文数	抽出比	新出 文数	累積 文数	累 積 抽出比
35	7	0.42			見える	7	3.55	+1	187	94.92
51	7	0.42	813	48.36	わき	7	3.55	0		
51	6	0.36			頭	6	3.05	+1	188	95.43
51	6	0.36			歩く	6	3.05	0		
51	6	0.36			いつ	5	2.54	0		
51	6	0.36			顔	5	2.54	0		
51	6	0.36	843	50.15	飛ぶ	5	2.54	0	188	95.43

表 7 たき火 (志賀直哉) における高頻度語と抽出文の関係

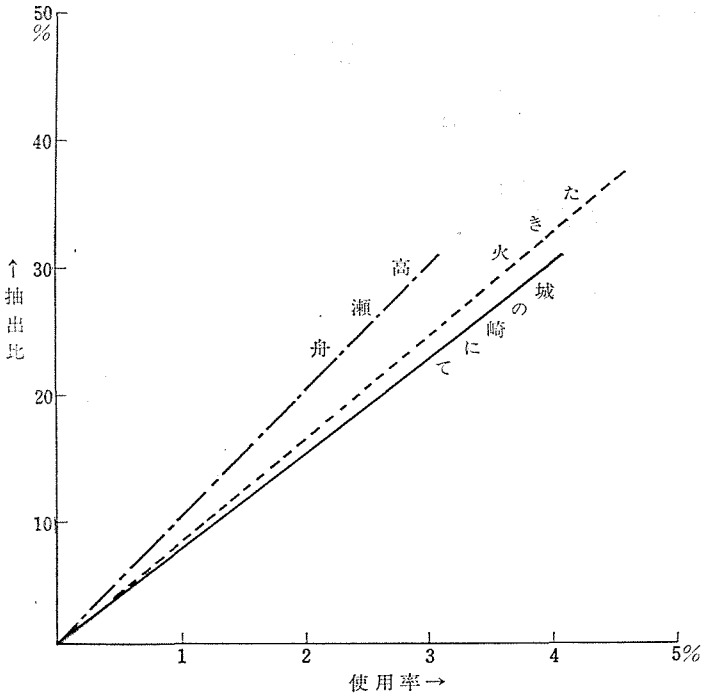
<延べ語数 2398 / 異り語数 802 / 総文数 278>

用 語					見 出 し	抽 出 文				
順位	度数	使用率	累積 度数	累 積 使用率		抽出 文数	抽出比	新出 文数	累積 文数	累 積 抽出比
1	109	4.55	109	4.55	さん	94	33.81		94	33.8
2	77	3.21	186	7.76	K	73	26.26	+1	95	34.2
3	70	2.92	256	10.68	いる	65	23.38	+35	130	46.8
4	62	2.59	318	13.26	言ふ	58	20.86	+17	147	52.9
5	54	2.25	372	15.51	する	46	16.55	+17	164	59.0
6	29	1.21	401	16.72	なる	28	10.07	+14	178	64.0
7	27	1.13	428	17.85	それ	27	9.71	+10	188	67.6
8	26	1.08	454	18.93	その	25	8.99	+6	194	69.8
9	25	1.04			良い	24	8.63	+4	198	71.2
9	25	1.04	504	21.02	S	24	8.63	0		
11	24	1.00	528	22.02	くる	23	8.27	+3	201	72.3
12	23	0.96	551	22.98	行く	21	7.55	+6	207	74.5
13	21	0.88	572	23.85	山	17	6.11	+3	210	75.5
14	20	0.83			妻	20	7.19	+4	214	77.0
14	20	0.83	612	25.52	ない	18	6.47	+3	217	78.1
16	19	0.79			事	19	6.83	+2	219	78.8
16	19	0.79	650	27.11	見る	19	6.83	+2	221	79.5
18	18	0.75			有る	18	6.47	+2	223	80.2
18	18	0.75	686	28.61	みんな	18	6.47	+7	230	82.7
20	17	0.71	703	29.32	中	17	6.12	+3	233	83.8
21	16	0.67			上	15	5.40	+2	235	84.5
21	16	0.67	735	30.65	自分	15	5.40	+4	239	86.0
23	12	0.50			聞く	12	4.32	0		
23	12	0.50			方(ハウ)	12	4.32	+2	241	86.7
23	12	0.50	771	32.15	見える	12	4.32	+1	242	87.1

用語					見出し	抽出文				
順位	度数	使用率	累積 度数	累積 使用率		抽出 文数	抽出比	新出 文数	累積 文数	累積 抽出比
		%		%		%			%	
26	11	0.46			この	11	3.96	0		
26	11	0.46			何(ナン)	11	3.96	+1	243	87.4
26	11	0.46	804	33.53	登る ○	10	3.60	0		
29	10	0.42			そして	10	3.60	+1	244	87.8
29	10	0.42			出す	10	3.60	0		
29	10	0.42			時(トキ)	10	3.60	0		
29	10	0.42			辺	10	3.60	0		
29	10	0.42			もの	10	3.60	0		
29	10	0.42	864	36.03	雪 ○	10	3.60	+1	245	88.1
35	9	0.38			小屋 ○	8	2.88	0		
35	9	0.38			さう	9	3.24	+1	246	88.5
35	9	0.38			しかし	9	3.24	0		
35	9	0.38			しまふ	9	3.24	0		
35	9	0.38			話 ○	8	2.88	0		
35	9	0.38			人(ヒト)	8	2.88	0		
35	9	0.38			舟 ○	9	3.24	+2	248	89.2
35	9	0.38	936	39.03	もう	8	2.88	+1	249	89.6
43	8	0.33			大きい	8	2.88	0		
43	8	0.33			おっかさん	7	2.52	0		
43	8	0.33			思ふ	8	2.88	0		
43	8	0.33			気持 ○	8	2.88	0		
43	8	0.33			頃	8	2.88	+1	250	89.9
43	8	0.33			先	8	2.88	0		
43	8	0.33			静か ○	8	2.88	+2	252	90.6
43	8	0.33			たき火 ○	8	2.88	0		
43	8	0.33			不思議 ○	8	2.88	0		
43	8	0.33			水 ○	7	2.52	0		
43	8	0.33	1024	42.70	森 ○	8	2.88	0	252	90.6
54	7	0.29			岸 ○	7	2.52	0	252	90.6
					……					
					……					
					……					

の関係を、グラフに表わしてみると、図3・図4のようになる。これらのグラフからわかるように、用語をめぐる数値の変動と、抽出文をめぐる数値の変動との間には、たいへん大きな開きがある。たとえば、図3の横軸と縦軸の目盛

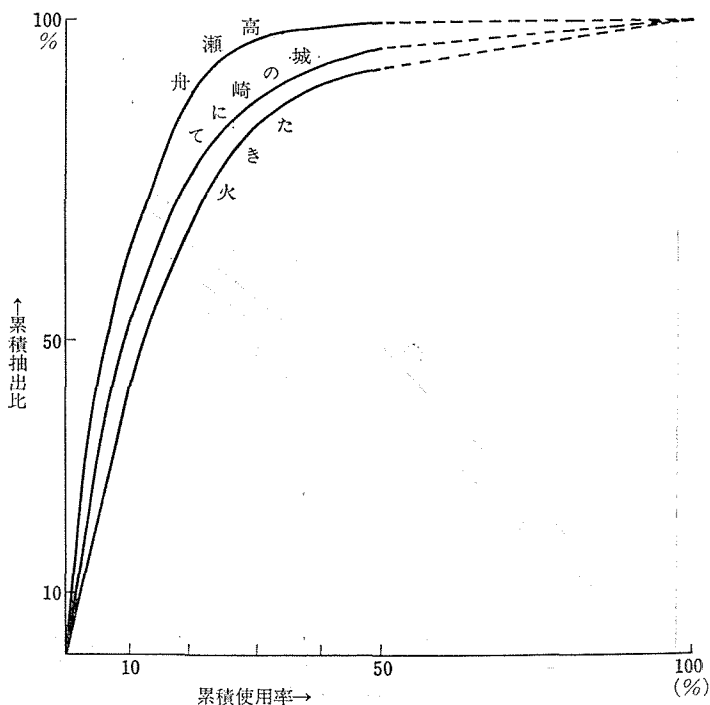
図 3



りからわかるように、文の「抽出比」の変動は、用語の「使用率」に比べると、その変動の振幅が圧倒的に大きい。また、図4のグラフからわかるように、文の「累積抽出比」も、用語の方の「累積使用率」に比べると、これまた圧倒的な伸び率で100パーセントに近づいていってしまう。たしかに、語の「使用率」と文の「抽出比」あるいは、語の「累積使用率」と文の「累積使用率」との間には、常に、一方が増加すれば他方も同様に増加し、一方が減少すれば他方も減少するというパラレルな関係があることは、これらのグラフからも読みとれるが、両者の変動領域・変動振幅が、著しく異なる点は重視すべきである。簡単に言えば、語の使用頻度の、ごく小さな変動が、文の抽出の面では、きわめて大きな変動となって現われやすいということであり、これは、ある場合には、抄録処理にとって、かなり致命的な影響を与えてしまう可能性がある。

このようなことが心配されるのは、言うまでもなく、キー・ワードの選定の基礎を、用語の頻度に求めたからにはほかならない。

図 4



自動抄録におけるキー・ワードは、結局は、文の抽出のキーとして働くものであることを考えれば、キー・ワードの選定も、その基礎を各語が抽出してくる文の数、あるいは抽出比・累積抽出比といったような抽出文をめぐるところに求めるべきではないだろうか。結果的には、多くの場合、用語の頻度にもとづくものと大差ないことになるかもしれないが、文を抽出する作業の基本的な手がかりが、便宜的に用語の頻度から割出されるよりは、実際にどれだけのセンテンスに影響する語がキー・ワードになっているかを常に把握しうる点だけでも、抽出文をめぐる数値にもとづく選定方法は、すぐれている。

また、実際の抄録処理の作業の上では、たいして支障にはならないかもしれないが、第3節表4に挙げた「蜘蛛の糸」の下位のキー・ワードたとえば「血・光・蓮……」のような抽出文ゼロとなるようなキー・ワードや、「たき火」の「森」のように、わずか一文しか抽出しえないキー・ワードなどが採用され

るようなことは、各語のセンテンス抽出能力を基礎にキー・ワードの選定を進めていけば、当然避けられるはずである。用語の使用頻度に基づいて選定されたキー・ワードの最大の欠点は、選定されたキー・ワードが、実際に原文のセンテンスをどのように抽出するか、言いかえれば、キー・ワードによって、原文のどの程度のセンテンスをおおい得るかということについての見通しを持っていないところにある。これを避けるためには、キー・ワード選定の基礎を、各語のセンテンス抽出能力におかなくてはならない。

たしかに、便宜的な方法として、あるいは近似的な手段として、語の使用頻度に基づくキー・ワードの選定方式は、処理の手続きも比較的簡単であるうえ、結果的には、かなりうまくセンテンス抽出能力を反映するものではある。しかし、たとえば、二つのキー・ワードの候補語のうち、使用頻度の高い方は原文の前半のセンテンス群に集中し、低いものは後半のセンテンス群に集中するというような場合、使用頻度のみからキー・ワードを決定すると、抽出文が原文の前半に偏ってしまう結果になる。実際の文章においては、話の進み具合に伴って、話題も変わってくるであろうし、別な話題がさしはさまれてくることも珍しくない。こうした、文章の変化・変質に対応しうのような抄録を目ざすとなると、すくなくとも、各キー・ワードの原文センテンスにおける分布状況ぐらいは、事前に把握しておかなくてはならない。第1節で紹介したように、水谷静夫氏が、章節ごとに抄出する方式を試みたのも、一つには、論旨の変化・変質を抄録文に反映させることを目ざしたものであろう。たしかに、処理のうで、こうしたキメ細かい手法を開発することも、もちろん重要であるが、その基礎となるキー・ワードも、そうした点を反映するように選定したいわけである。

以上のように考えると、一般の文章の自動抄録を進めるためのキー・ワードは、最低の条件として、原文のセンテンスとの関連において選定されていないということになる。

それでは、その場合、何を手掛りにして、どのようにキー・ワードの選定を進めるべきかとなると、まだ、はっきりと答えうる段階ではないが、少なくとも、各語のセンテンス抽出能力をはかる測度として、さきに挙げたような、抽

出比・新出文数・累積抽出比などは、まず採り上げるべきものではないかと思う。さらに欲をいえば、各語が、原文のどのあたりのセンテンスをおおっているか、すなわち、前半のセンテンスを主に抽出するか、後半のセンテンスを集中的に抽出するかといったことも、合わせて考えるべきであろう。コンピュータを使えば、キー・ワード選定の手掛りを、このようなところに求めることは、それほど困難なことではない。

しかし、そのためには、さきに表5・表6・表7に掲げたような基本的なデータを各分野の文章について求めることが、まず第一歩である。それに基づいて、センテンスの抽出能力を手がかりとして、原文のセンテンスとの関連において、キー・ワードを選定していく方法を確立することが、今後の課題ではないかと考えるのである。

8. キー・ワードの長さ

自動抄録のキー・ワードをめぐっては、キー・ワード自体の長さ、すなわちキー・ワードとしては、どのような長さの言語単位を採用すべきかということも、ゆるがせにできない問題である。

今回の実験では、ある程度、原文の性格を考慮して、キー・ワードの選定過程において、「高瀬一舟」「高瀬一川」を一語とするといった措置を、一部試みではみたが、大体において、かなり短い単位を採用した。それは、一つには「無性格語」のリストを、従来の語彙調査の結果を使って作成したため、これら(注6)の語彙調査において採用されている単位(β単位)から、あまり大きくはずれられないわけにはいかなかったという理由による。

しかし、考えてみると、自動抄録のキー・ワードは、本来、原文の文章を特徴づけるような、「いかにも、その文章らしい語」を選ぶべきものであるから、あまり短い単位よりは、ある程度長い単位の方が、当然、特徴が出てきやすい。たとえば「高瀬舟」のキー・ワードを選ぶ以上、「第一殺し」や「島一送り」などの単語は、「弟／殺し」「島／送り」と分割してしまうよりは、全体で一語として扱った方が、「高瀬舟」らしさをそこなわないというわけである。この点について、かつて木村繁氏は、論文「層別特徴語の判別」(注8)において、新

聞語彙調査の各層の特徴語を分析し、長い単位においてはじめて特徴語となるものがあることを指摘している。

自動抄録においても、たとえば表4に挙げた「蜘蛛の糸」の場合、「蜘蛛」「血」などは、もし、これをキー・ワードとして単純に高頻度語からセンテンス抽出を行なったとすると、これらの語が抽出してくるセンテンス（新出文）は、きわめて少ない。「蜘蛛」は、頻度順位第5位のキー・ワードとランクされるものでありながら、わずか1文しか抽出してこない。「血」にいたっては、度数7でありながら、表4に示す通り、これが抽出するセンテンスはゼロである。これは、原文において、「蜘蛛」が、ほとんど「蜘蛛-の-糸」の形で使われていたためであり、「血」の場合は、そのすべてが「血の池」の形で使われていた結果である。したがって「蜘蛛の糸」を「蜘蛛/の/糸」と単位切りしてしまって「蜘蛛」と「糸」とを重複してキー・ワードにしてみても、「血の池」を「血/の/池」と分割して「血」「池」を別々にキー・ワードとしてみても、センテンス抽出のさいの実際的な効果は、ほとんどなかったわけである。これらは、「蜘蛛」か「糸」かのどちらかを、あるいは「血」と「池」のいずれか一方をキー・ワードとして採用すれば十分だったともいえるが、キー・ワードの「特徴語」としての性格を重視するならば、むしろ「蜘蛛の糸」「血の池」の形でキー・ワードとした方が適切なのではなからうか。同じような理由によって、やはり「蜘蛛の糸」の「お釈迦さま」を「お/釈迦/さま」と切った形で扱い、「釈迦」をキー・ワードとするのも、原文において「釈迦」がすべて「お釈迦さま」の形で使われている以上、あまり意味がない。

このように考えてくると、原文を特徴づけ、かつセンテンスの抄出の面でも有効なキー・ワードを選定するためには、各々の語が原文にどのような形で出てくるか、あるいは、どんな語と関連しあって出てくるかを事前に把握しておかなくてはならないということになる。それと同時に、キー・ワードには、ある程度長い語形、長い単位を認めた方が有効だということも浮かびあがってくる。

しかし、キー・ワード選定のプロセスの、どの段階において、長い単位を採用すべきかは、一考を要する。なぜなら、最初から長い単位を採用するとなる

と、無性格語のリストも長い単位で作らざるをえなくなり、必然的に無性格語リストが、かなり大きなものになってしまう。したがって、最初は、短い単位で、キー・ワード候補を選定し、それらについて、原文の中における実際の現われ方や語形を確かめた上で、ものによっては長い単位を採用していくというような方法も一つのやり方ではなかろうか。操作上の細かい点については、種々検討すべき余地があるが、いずれにしても、キー・ワードの語長は、語彙調査などの場合と異なり、必ずしも等質でなければならない理由はない。むしろ、原文の特性をよく反映し、センテンス抽出に十分な効果をあげるキー・ワードを選ぶためには、原文に即した長さを求めるべきではないかと思うのである。この点を、特に重視するならば無性格語との間で、言語単位を揃える必要はないともいえよう。

注8) 国立国語研究所報告34「電子計算機による国語研究(Ⅱ)」所収

9. 無性格語の選定

最後に、キー・ワード選定の基本的なプロセスにおいて、重要な役割を果す「無性格語」について考えてみたい。

今回の実験では、無性格語は、主として大量語彙調査の調査結果にもとづいて決定したが、やはり、もともなった語彙調査が、すべて雑誌を対象としたものであったため^(注6)、それによる偏りを免れえなかったようである。その最も顕著なところは、話しことば的な色彩をもつ基本語が、ほとんど「無性格語リスト(表3)」に入らなかった点である。たとえば、「ございます」の「ござる」や「いたします」の「いたす」や、「ください」の「くださる」など、話しことばにおいては、きわめて高い頻度で出てくる語が、ほとんど拾われていない。また、「わたくし」はあるが「わたし」「あなた」といった人称代名詞がないとか、「はい」「ええ」といった類の感動詞がまったく収容されていないのも、同じ理由からであろう。したがって、このような「無性格語リスト」にたよりきって、たとえば小説のように対話文体の多いものの抄録を行なうわけにはいかないということになる。今回の実験で、途中で一度キー・ワード候補をアウト・プットしたのも、一つには、「無性格語リスト」が、まだ不十分

だったからである。

自動抄録の場合の「無性格語」というものは、結局「キー・ワードにする必要のない語」であるから、たしかに、大量語彙調査で高頻度を占めるような「ありふれた語」が、その一要素であることは疑いない。しかし、自動抄録が、実際には、一つ一つ文章を対象にして行なわれることを考えると、たとえば小説の文章を対象とする場合には、小説類に共通して出てくるような「ありふれた語」をプラスした方が、より有効だということになる。小説類の場合なら、すくなくとも対話的な性格をもつ語を補うことが、ぜひとも必要である。また、科学技術論文の抄録を行なう場合なら、同じ分野の論文一般に出てくるような語は加えておくべきであろう。これは、いうなれば、文章のジャンル別の基本語である。文章のジャンルとして、どのようなものをたてていくとか、各ジャンル特有な基本語をいかして選ぶかとかいったような、さまざまな問題が残っているが、たとえば、小説類については、この稿に掲げた表1・表2・表5・表6・表7の類の作品の語彙調査結果をつき合わせていけば、自づから、小説という文章ジャンルに共通に出てくる「ありふれた語」が浮かびあがってくる（その場合の手法としては、水谷氏が、雑誌九十種の語彙調査において採用した、用語の層別の分布から「語の基本度の表」^(注9)を作成した方法などが、よい参考になる）。

さらにつけ加えれば、上記のような、基本的な語や共通性のある語ばかりでなく、最初からキー・ワードとなり得ない語群、さきに4・1に挙げたような「上・下・右・左・前・うしろ・あと・さき・間」など「関係概念を表わす名詞」や代名詞・指示語の類、および副詞・連体詞・接続詞・感動詞などは、「無性格語リスト」に、はじめから入れておいた方が賢明である（これらの選定には、林大氏の作成した「分類語彙表」^(注10)や、新聞の語彙調査のさいに中野洋氏の作成した「品詞別語彙表」^(注11)などが役立つ）。

以上のようなものを収容した、一応、理想的な「無性格語リスト」が、整理されていれば、今回の実験において、途中で一度キー・ワード候補を出力して、たとえば「わたし・あの・ただ・あなた」の類（表2の△印）を削除する操作は、ほぼ不要になる。そうしてこそ、はじめて、一貫的な自動抄録処理

が、かなり高い精度で可能になるわけである。

従来の自動抄録論においては、どちらかという、キー・ワードをめぐる問題や、それによるセンテンス抽出の手法の問題に重点がおかれ、「無性格語」は、あまり重視されてこなかった。わずかに、「語の基本度の表」の適用を提唱した水谷氏の論がある程度である。^(注11)しかし、「無性格語」はキー・ワード選定の前提となるものであり、これが的確に選ばれているかどうかは、キー・ワードの質と、そのセンテンス抽出における有効性を決定する点で、きわめて重要なはずである。自動抄録処理のプロセスからみて、「無性格語リスト」との照合によって、キー・ワードとして不適格な語が確実にふるい落とされるか否かは、その後の処理の効率と、抄録文の出来具合の良否に大きな影響を与えるものである。

今回の実験では、無性格語リストについては、水谷氏の提案以上の新味は、ほとんど加ええなかったが、自動抄録処理における無性格語の重要性と、その選定のあり方に触れて、この論を終わることにする。

注9) 国立国語研究所報告25「現代雑誌九十種の用語用字(第三分冊)」の「1・語の基本度(水谷静夫)」

注10) 国立国語研究所資料集6「分類語彙表」

注11) 国立国語研究所報告38「電子計算機による新聞の語彙調査(Ⅱ)」の「IV.品詞別度数順短単位表」