文節の係り受け距離の統計分析

金 明哲

In this paper, the author tried statistical analysis of phrase to phrase modification distance in the novels. The statistical analysis includes in two parts, one is for the relationship between the phrase to phrase modification distance and writer's individuality, and another is for the relationship between the phrase to phrase modification distance and commas. The analysis shows that the writer's individuality is not reflected in the distributions of the phrase to phrase modification distance. The distributions are usually subordinate to L type distribution. But the distributions are not subordinate to the same distribution, when the distributions of the phrase to phrase modification distance were divided by range. There is no certain relationship between the distribution of the phrase to phrase modification distance and commas. As for the statistical method, the average value of distribution distance (χ^2 distance), *AIC* (Akaike Information Criterion) and chi-square measurement are used in the research.

1．はじめに

文節の係り受け規則は日本語の機械処理によく用いられている．古田[10]は二文節間の係り受け関係を基礎として日本語の構文解析を行い，武石・林[6]は係り受け構造に基づいた日本語の複文の分割方法を提案した．任ら[7]は機械翻訳における曖昧さの解消を原言語の係り受け関係に着目し研究を試みた．黒橋・長尾[4]は「単純な係り受け規則を適用するだけで高度な構文解析が可能である」という結果を得た．このような日本語の機械処理に関する研究に伴い，日本語における文節間の係り受け関係の統計的特性を明らかにすることには重要な意味がある．日本語における文節間の係り受け関係の統計的性質に関する研究としてはMaruyama, H. and Ogino, S.[8]がある．Maruyama, H. and Ogino, Sは新聞記事文（8352文節）を用いて分析を行い，文節の係り受け関係はレンジに関係なくZipf法則に従うという見解を示した．このような性質が普遍性を持っているか否かは，今後の日本語の研究において極めて重要である．しかし，他のジャンルの文でもこのような統計的性質を持っているか否かに関しては明らかにされていない．また，文節間の係り受け距離が遠くなると読点を打つと言われているが，係り受け距離と読点の打ち方との関係に関する統計分析には先例がない．

本稿では，このような未解決の問題を視野に入れ，3人の小説文（46786文節）を用いて書き手と係り受け距離の分布，係り受け距離

JIN Mingzhe 札幌学院大学社会情報学部

と読点との関係について行った計量分析の方法およびその結果を述べる．計量分析に用いた文章は，井上靖，三島由紀夫，中島敦の短篇小説である．統計分析に用いる情報の抽出のため，それらの文章を計算機可読なようにデータベース化した．文章毎に抽出された情報の安定性を見るため，これらの小説の中で比較的長い小説はいくつかに分割して用いた．例えば，井上の「恋と死と波と」は二つに，中島の「弟子」は三つに，「李陵」は四つに分割して用いることにした．ただし，文章に含まれている会話文及び引用文は除いた．表1に用いた文章と発表年などを示した．なお表2に作成した3人のデータベースの一部分を示す．

データベースは，分析に用いる文章をOCR（光学読み取り機器）で入力し，読み取りの誤りを訂正し，単語毎に品詞コード，文節の境界記号，文節の係り受け番号などを入力して作成した（手作業）．

表1　計量分析に用いた文章

著　者	文　章　名	文章の記号	単語数	文節数	文数	出版社	発表の年
井上靖	結婚記念日	I 1	4749	2543	195	角川文庫	1951
	石庭	I 2	4796	2544	237	同上	1950
	死と恋と波と（前半）	I 3	4683	2551	234	同上	1950
	死と恋と波と（後半）	I 4	4386	2359	241	同上	同上
	帽子	I 5	3724	1902	177	新潮文庫	1973
	魔法壜	I 6	3624	1886	138	同上	同上
	滝へ降りる道	I 7	3727	1934	164	同上	1952
	晩夏	I 8	4269	2230	222	同上	同上
三島由紀夫	遠乗会	M 1	4984	2720	287	新潮文庫	1951
	卵	M 2	4004	2155	178	同上	1955
	詩を書く少年	M 3	4502	2390	247	同上	1955
	海と夕焼	M 4	3359	1890	201	同上	1955
中島敦	山月記	L 1	3226	1743	179	新潮文庫	1942
	名人伝	L 2	3202	1741	155	同上	1942
	弟子（前の1/3）	L 3	4078	2210	212	同上	1943
	弟子（中の1/3）	L 4	4092	2201	227	同上	同上
	弟子（後の1/3）	L 5	3727	2044	227	同上	同上
	李陵（前の1/4）	L 6	4563	2481	207	同上	1944
	李陵（中の1/4）	L 4	4561	2446	203	同上	同上
	李陵（中の1/4）	L 8	4638	2482	219	同上	同上
	李陵（後の1/4）	L 9	4458	2334	185	同上	同上

表2　データベースの例

(7)／私（M）は（J）(7)／小学校（M）を（J）(4)／郷里（M）(5)／伊豆（M）の（J）(6)／祖母（M）の（J）(7)／許（M）で（J）(8)／過し（D）た（Z）．

(2)／父（M）は（J）(27)／軍医（M）で（Z），(4)／当時（M）(5)／聯隊（M）の（J）(6)／ある（R）(6)／地方（M）の（J）(9)／小都市（M）を（J）(9)／転々と（F）(10)／し（D）と（J）(27)／おり（D），(11)／子供（M）を（J）(13)／自分（M）の（J）(14)／手許（M）に（J）(27)／置く（D）と（J），(16)／何回（M）も（J）(17)／転校させ（D）なけれ（Z）ば（J）(23)／なら（D）なかっ（Z）た（Z）ので（J），(19)／そう（F）(20)／し（D）た（Z）(23)／こと（M）から（J）(23)／私（M）を（J）(23)／郷里（M）に（J）(24)／置く（D）(25)／気（M）に（J）(26)／なっ（D）た（Z）(27)／もの（M）らしいかっ（Z）た（Z）．

2．文の構造の数値化

文の構造を捉えるには，主語，述語などの成分で文の構造を表すことが一般的である．しかし，本研究では文の構造を数値で表すため，文の中の成分を二つの文節の係り受けの関係で表すことにする．具体的な規則を下記に示す．

1．文の中の文節に頭から番号を付け，その文節が何番目の文節に係るか，その数値をその文節に与える．
2．最終の述語は，次の番号を持つ「終止」に係るものとする．
3．接続詞のうち，「並立，逆説，同時進行」などを表すもの(そして，また，すると等)は「終止」に係るものとする．
4．接続語のうち，「仮定，理由」などの条件を表すもの（だから，故に等）は述語に係るものとする．
5．感動詞は「終止」に係るものとする．
6．主語や連体修飾語などが同時に二つ以上の文節に係る場合には，最初の要素に係るものとる．
7．並立する文節の係り受け先は，並立する文節の最後のものの係り受け先とする．
8．並立する文節全体に係る文節は，並立する文節の最初のものに係るものとする．
9．提示の文節で，その後にそれを指す文節がある場合には，提示の文節は，指す文節の係り先に係るものとする．

例えば，文
山の中の暗い夜には星が砂を撒いたように見える．
を上述の規則で，文節ごとに文節の番号[i]と係り受け先の文節の番号（j）をつけると次のようになる．

[1]山の(2)/[2]中の(4)/[3]暗い(4)/[4]夜には(8)/[5]星が(8)/[6]砂を(7)/[7]撒いたように(8)/[8]見える(9)/[9][(注1)]

この例文の係り受け関係は図1のように表すことができる．

文の中のi番目の文節がj番目の文節に係る時，$j-i$をi番目の文節の係り受け距離と定義し，以下では単に距離と呼ぶことにする．上記の例文の文節間の関係を距離別に分けて表すと以下のようになる．

距離1　山の ——→ 中の
　　暗い ——→ 夜には
　　砂を ——→ 撒いたように
　　撒いたように ——→ 見える
　　見える ——→ .
距離2　中の ——→ 夜には
距離3　星が ——→ 見える
距離4　夜には ——→ 見える

3．係り受け距離の分布

3.1　係り受け距離の分布と書き手の個性

本節では文の構造と書き手との関係について分析を行なう．文の構造を係り受け距離で考えた場合，短い距離の修飾関係で文を書くと文が易しく理解しやすいが，長い距離を持つ修飾関係を多く用いると文が難しくなると考えられる．書き手によって文章の構造にこのような差があるであろうか．表3に文章における係り受け距離が1から10までおよび11以上，合計11カテゴリーに分けた場合の文節の使用頻度・相対頻度を示した．上の行は使用頻度で，下の行は相対頻度である（以

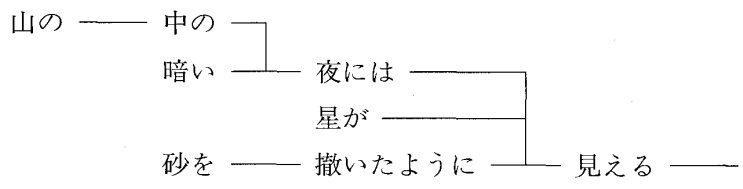

図1　文の構造図

表3 文節の係り受け距離の分布

著者	文章	1	2	3	4	5	6	7	8	9	10	11以上
井上	I 1	1579	282	177	133	93	61	28	25	23	27	115
		0.621	0.111	0.070	0.052	0.037	0.024	0.011	0.010	0.009	0.011	0.045
	I 2	1621	332	185	82	71	54	44	29	21	16	89
		0.637	0.131	0.073	0.032	0.028	0.021	0.017	0.011	0.008	0.006	0.035
	I 3	1620	304	155	112	86	59	36	34	23	19	103
		0.635	0.119	0.061	0.044	0.034	0.023	0.014	0.013	0.009	0.007	0.040
	I 4	1538	295	145	91	62	50	42	27	21	16	72
		0.652	0.125	0.061	0.039	0.026	0.021	0.018	0.011	0.009	0.007	0.031
	I 5	1269	172	136	92	53	43	24	25	21	11	56
		0.667	0.090	0.072	0.048	0.028	0.023	0.013	0.013	0.011	0.006	0.029
	I 6	1246	183	114	79	53	39	41	23	17	18	73
		0.661	0.097	0.060	0.042	0.028	0.021	0.022	0.012	0.009	0.010	0.039
	I 7	1247	235	120	79	53	32	29	21	28	12	78
		0.645	0.122	0.062	0.041	0.027	0.017	0.015	0.011	0.014	0.006	0.040
	I 8	1445	248	145	101	70	37	35	44	20	23	62
		0.648	0.111	0.065	0.045	0.031	0.017	0.016	0.020	0.009	0.010	0.028
三島	M 1	1777	332	172	124	76	51	45	30	27	13	73
		0.653	0.122	0.063	0.046	0.028	0.019	0.017	0.011	0.010	0.005	0.027
	M 2	1349	253	143	89	68	58	43	26	18	21	87
		0.626	0.117	0.066	0.041	0.032	0.027	0.020	0.012	0.008	0.010	0.040
	M 3	1542	300	157	110	71	46	32	18	24	19	71
		0.645	0.126	0.066	0.046	0.030	0.019	0.013	0.008	0.010	0.008	0.030
	M 4	1244	227	113	87	48	50	31	24	14	12	40
		0.658	0.120	0.060	0.046	0.025	0.026	0.016	0.013	0.007	0.006	0.021
中島	N 1	1112	202	123	73	63	37	34	19	12	7	61
		0.638	0.116	0.071	0.042	0.036	0.021	0.020	0.011	0.007	0.004	0.035
	N 2	1079	212	133	87	62	42	26	14	22	21	43
		0.620	0.122	0.076	0.050	0.036	0.024	0.015	0.008	0.013	0.012	0.025
	N 3	1436	275	149	100	59	52	32	26	23	12	46
		0.650	0.124	0.067	0.045	0.027	0.024	0.014	0.012	0.010	0.005	0.021
	N 4	1439	272	141	97	69	43	35	24	15	16	50
		0.654	0.124	0.064	0.044	0.031	0.020	0.016	0.011	0.007	0.007	0.023
	N 5	1361	226	148	82	65	33	37	17	22	6	47
		0.666	0.111	0.072	0.040	0.032	0.016	0.018	0.008	0.011	0.003	0.023
	N 6	1568	320	163	109	81	49	47	34	17	15	78
		0.632	0.129	0.066	0.044	0.033	0.020	0.019	0.014	0.007	0.006	0.031
	N 7	1610	239	165	98	58	53	45	30	19	24	105
		0.658	0.098	0.067	0.040	0.024	0.022	0.018	0.012	0.008	0.010	0.043
	N 8	1585	281	180	114	72	53	38	29	22	28	80
		0.639	0.113	0.073	0.046	0.029	0.021	0.015	0.012	0.009	0.011	0.032
	N 9	1493	276	180	100	67	50	31	30	15	16	76
		0.640	0.118	0.077	0.043	0.029	0.021	0.013	0.013	0.006	0.007	0.033

下係り受け距離の分布と呼ぶことにする.)

文章 i における距離 j の文節の使用頻度を x_{ij} で表すと，3人の21文章における11カテゴリーに分けた係り受け距離の使用頻度はマトリックス

$$X_{21\times 11}=[x_{ij}]$$

で表記でき，相対頻度のマトリックスは

$$P_{21\times 11}=[p_{ij}] \qquad p_{ij}=\frac{x_{ij}}{\sum_{v=1}^{11} x_{iv}}$$

で表記できる．図2に $P_{21\times 11}$ のプロットを示した．図2で，係り受け距離の分布はかなりよく一致することがわかる．分布間の差を数値で考察するため分布間の距離を求めてみた．本節では χ^2 距離を用いた．

$$\chi^2(i,l)=\frac{x_{i\cdot}\,x_{l\cdot}}{x_{i\cdot}+x_{l\cdot}}\sum_{j=1}^{11}\frac{(p_{ij}-p_{lj})^2}{p_{ij}+p_{lj}}$$

$$x_{i\cdot}=\sum_{j=1}^{11} x_{ij}, \qquad x_{l}=\sum_{j=1}^{11} x_{lj},$$

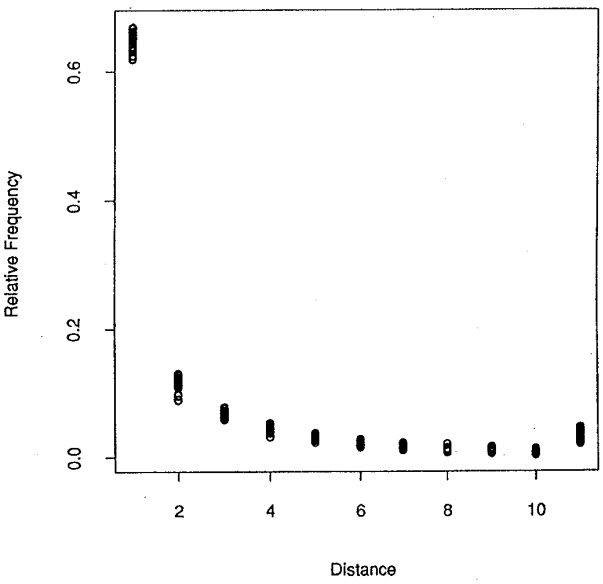

図2 係り受け距離の分布のプロット

$$p_{ij}=\frac{x_{ij}}{\sum_{v=1}^{11} x_{iv}}, \qquad p_{lj}=\frac{x_{lj}}{\sum_{v=1}^{11} x_{iv}}$$

一般的には，係り受け距離の分布に書き手の個性が現われるとすると，同一著者の各文章における係り受け距離の分布間の距離（以下群内距離と呼ぶ）の平均値は，異なる著者の各文章における係り受け距離の分布間の距離（以下群間距離と呼ぶ）の平均値より小さいことになる．著者 k の i 番目の文章を k_i，著者 k の j 番目の文章を k_j と，群内の距離の平均は，

$$\overline{d(k)}=\frac{\sum_{k_i=k_1}^{k_n}\sum_{k_j=k_i+1}^{k_n}\chi^2(k_i,k_j)}{k_n(k_n-1)/2},$$

著者 h の j 番目の文章を h_j と表記すると，群間の距離の平均は，

$$\overline{d(k,h)}=\frac{\sum_{k_i=k_1}^{k_n}\sum_{h_j=h_1}^{h_m}\chi^2(k_i,h_j)}{k_n\,h_m}$$

で求められる，表4に上式で求めた群内，群間の距離の平均値を示した．

群内の距離と群間の距離とを比べてみると，井上，中島の群内の距離は最小の群間の距離より大きい．χ^2 距離がもっとも大きいのは井上の群内の距離で8.210である．これは $\chi^2(df=10,\ 5\%)=18.31$ より大きく下回る．以上の結果から，係り受け距離の分布は

表4 文節の係り受け距離の分布の χ^2 距離の平均値

著者	群内	群		間	最小の群間
井上	8.210		7.454	7.990	7.454
三島	6.276	7.454		6.148	6.148
中島	7.131	7.988	6.148		6.148

文章および書き手に関係なく同じ分布に従うと考えられる．つまり，係り受け距離には書き手の個性が見られないということがデータから読み取られる．

3.2 レンジ毎の書き手の差

文節の係り受け距離の分布は文節の位置と関係があるか否かについて検証を行うため，本節と次の節では，レンジ(注2)毎の係り受け距離の分布について分析を行なう．いま，文節は単位とした長さ n の文を文頭から順番に文節に番号を付ける．

文節(1)/文節(2)/……/文節($j-1$)/文節(j)/文節($j+1$)/……/文節(n)．

文の中の文節(j)は右側の $n-j$ 個の文節の中の一つの文節を修飾する．ここでは $r=n-j$ をレンジと呼ぶことにする．論文の冗長を避けるため，レンジ毎に分けた場合の同一著者における係り受け距離の分布間の差に関する計量分析は割愛し，同一著者による各文章におけるレンジ別の係り受け距離を一つにまとめ分析を進める．

表5に著者別のレンジ毎の係り受け距離の分布を示した．上の行は使用頻度で，下の行は相対頻度である．

レンジ i における係り受け距離 v の使用頻度を x_{iv}，相対頻度を p_{iv} に表すと

$$p_{iv}=\frac{x_{iv}}{\sum_{j=1}^{r} x_{ij}} \qquad \sum_{v=1}^{r} p_{iv}=1$$

表5の3人のレンジ別の係り受け距離の分布に差があるかどうかを検証してみることにする．著者 k のレンジ r での係り受け距離 v の使用頻度 x_{kv} で表すと，3人のレンジ r で

表5 レンジ毎の係り受け距離の分布

距離	1	2	3	4	5	6	7	8	9	10	$x>10$
レンジ			井					上			
3	1158	357	59								
	0.736	0.227	0.037								
4	897	275	227	91							
	0.602	0.185	0.152	0.061							
5	758	174	206	138	114						
	0.545	0.125	0.148	0.099	0.082						
6	729	143	102	114	83	107					
	0.570	0.112	0.080	0.089	0.066	0.084					
7	625	138	74	72	94	61	108				
	0.533	0.118	0.063	0.061	0.080	0.052	0.092				
8	602	110	54	54	45	50	43	109			
	0.564	0.103	0.051	0.051	0.042	0.047	0.040	0.074			
9	538	108	56	36	35	30	41	29	83		
	0.563	0.113	0.059	0.038	0.037	0.031	0.043	0.030	0.087		
10	450	115	56	28	16	21	20	35	27	72	
	0.536	0.137	0.067	0.033	0.019	0.025	0.024	0.042	0.032	0.086	
11以上	417	88	50	33	12	20	10	7	26	24	57
	0.560	0.118	0.067	0.044	0.016	0.027	0.013	0.009	0.035	0.032	0.077
			三					島			
3	645	222	29								
	0.720	0.248	0.032								
4	506	146	133								
	0.609	0.176	0.160	0.055							
5	418	115	98	95	45						
	0.542	0.149	0.127	0.123	0.058						
6	355	103	59	73	55	43					
	0.516	0.150	0.086	0.106	0.080	0.062					
7	338	59	41	24	43	50	49				
	0.560	0.098	0.068	0.040	0.071	0.083	0.081				
8	275	58	39	25	20	36	24	38			
	0.534	0.113	0.076	0.049	0.039	0.070	0.047	0.074			
9	234	81	30	23	16	6	18	18	35		
	0.508	0.176	0.065	0.050	0.035	0.013	0.039	0.039	0.076		
10	214	52	24	22	11	12	8	8	14	27	
	0.546	0.133	0.061	0.056	0.028	0.031	0.020	0.020	0.036	0.069	
11以上	174	39	22	11	9	7	10	8	12	14	32
	0.515	0.115	0.065	0.033	0.027	0.021	0.030	0.024	0.036	0.041	0.095
			中					島			
3	1315	358	92								
	0.745	0.203	0.052								
4	1015	290	280	115							
	0.597	0.171	0.165	0.068							
5	850	213	202	188	135						
	0.535	0.134	0.127	0.118	0.085						
6	793	172	121	126	124	121					
	0.544	0.118	0.083	0.086	0.085	0.083					
7	738	146	82	93	77	71	131				
	0.552	0.109	0.061	0.070	0.058	0.053	0.098				
8	676	143	67	42	56	49	47	111			
	0.568	0.120	0.056	0.035	0.047	0.041	0.039	0.093			
9	575	122	79	37	30	33	51	28	106		
	0.542	0.115	0.074	0.035	0.028	0.031	0.048	0.026	0.100		
10	512	125	67	38	19	18	20	23	26	90	
	0.546	0.133	0.071	0.041	0.020	0.019	0.021	0.025	0.028	0.096	
11以上	453	103	64	37	13	14	13	17	13	12	71
	0.559	0.127	0.079	0.046	0.016	0.017	0.016	0.021	0.016	0.015	0.088

の係り受け距離は

$$R_{3\times r}=[x_{kv}]$$

で表される．分布の同一性を判断する数理的方法としては頻度マトリックス $R_{3\times r}$ における AIC，χ^2 の統計量を用いる方法がよく知られている．3人のレンジ毎の係り受け距離の分布が同じかどうかを判定するための AIC 統計量[6]は，$R_{3\times r}$ に対し式

$$AIC=-2\sum_{k=1}^{3}\sum_{v=1}^{r}x_{kv}\log\frac{x_{kv}}{e_{kv}}+2(3-1)(r-1)$$

を用いて求める．

もし　$AIC>0$ ならば分布は同じ，

もし　$AIC<0$ ならば分布は異なる，

もし　$AIC=0$ ならば判断不能，

と判定する[6]．ただし，

$$N=\sum_{k=1}^{3}\sum_{v=1}^{r}x_{kv},\qquad x_{k\cdot}=\sum_{v=1}^{r}x_{kv},$$

$$x_{\cdot v}=\sum_{k=1}^{3}x_{kv},\qquad e_{kv}=\frac{x_{k\cdot}\,x_{\cdot v}}{N}$$

である．χ^2 統計量としては，

$$\chi^2=\sum_{k=1}^{3}\sum_{v=1}^{r}\frac{(x_{kv}-e_{kv})^2}{e_{kv}}$$

$$G^2=2\sum_{k=1}^{3}\sum_{v=1}^{r}x_{kv}\log\frac{x_{kv}}{e_{kv}}$$

を用いた．表6に求めた AIC，χ^2，G^2 の統計量を示した．

表6から分かるように AIC 統計量を用いた場合は，レンジ3を除くと3人の係り受け距離の分布には差がないという結果が得られた．χ^2，G^2 統計量を用いた場合は，レンジ3，7，11以上を除くと3人の係り受け距離の分布には有意水準5％では差があると認められない．したがって，レンジ毎における係り受け距離の分布には書き手の個性が見られないと利断する．

3.3 レンジ間の差

表5をみた限りでは，同じ著者のレンジ毎の係り受け距離の分布には差があるように見られる．例えば，距離1の出現率はレンジ3の場合は約0.72～0.75であるのに対してレンジ11以上の場合は距離1の出現率は約0.52～0.56である．表5のデータを用いて統計的に分布の同一性を検証してみることにする．表5ではレンジ毎の距離の分布には距離に関するカテゴリー数が一致しない．したがって，AIC，χ^2，G^2 統計量を求めるためには，同数のカテゴリーを切り取って用いることにする．本稿では，指定されたレンジ r に対し，r より大きいレンジより構成された表（集合）の中から距離のカテゴリーを $r-1$ まで切り取って用いることにする．距離のカテゴリーを $r-1$ まで切り取って用いるのは，レンジ r で距離 i が $i<r$ の場合，p_r は単調減少関数であるが，$r=i$ になるとこの規則を満たさない場合があるからである．このような一連の作業を関係代数式で表現すると以下のようになる．

$$\pi_{1\sim(c-1)}(\sigma_{r>=c}(\mathrm{Table}))$$

式のなかの $Table$ はレンジ毎に分けた係り受け距離の分布表，r はレンジ，c はレンジの特定値である．π，σ はそれぞれ関係代数の射影，選択を表す[9]．たとえば，表5における井上のレンジ9以上（$c=9$）の場合は

$$\pi_{1\sim8}(\sigma_{r>=9}(\mathrm{Inoue}))$$

である．前節ですでにレンジ毎の係り受け距離分布には書き手の個性が見られないこと

表6　3人著者のレンジ毎の係り受け距離の分布の AIC，χ^2，G^2 統計量

レンジ	AIC	χ^2	G^2	$\chi^2_{0.05}$
3	**-5.18**	**13.23**	**13.18**	9.49
4	8.76	3.23	3.24	12.59
5	2.01	13.64	13.99	15.51
6	3.26	16.82	16.75	18.31
7	1.95	**22.09**	**22.05**	21.03
8	8.69	19.83	19.32	23.68
9	4.98	**27.23**	**27.02**	26.30
10	20.48	15.83	15.52	28.87
11以上	9.62	29.92	30.38	31.41

が確認されたため，本節では3人の係り受け距離の分布を一つにまとめ分析を進める．表7に $c=3, 4, 5, 6, 7, 8, 9, 10$ における任意の二つの分布の間の AIC, χ^2, G^2 統計量に基づいた分布が同じと認定される回数などを示した．分布が同じか否かの比較はそれぞれの c の値においてすべての分布間の組み合わせについて比較を行った．合計120回の比較のなか，AIC 統計量を用いた場合は，分布が同じであると認められる回数は55回（約46%）で，χ^2, G^2 統計量を用いた場合は，有意水準5%で分布が同じであると認められる回数はそれぞれ68回（約57%），66回（約55%）である．いずれの場合でもレンジに関係なく分布が同じであると認められるのは60%にもならない．分布のカテゴリー数が極端に少ない $c=3$ の場合を除くと，分布が同じであると認められるのは50%にもならない．したがって，文節の係り受け距離の分布がレンジとは関係がないと結論を出すには無理がある．視覚的に考察を行うため，図3に分析に用いたすべての文章におけるレンジ毎に分けた場合の文節の係り受け距離の分布を示した．

レンジの値が大きい時，係り受け距離がレンジの値と等しい場合の出現率 p_{ir} は，距離3～$r-1$ の出現率 p_{i3}～p_{ir-1} より大きいことが興味深い（$p_{ij}<p_{ir}$, $2<j<r$）．この結果は機械的にヒューリスティクスな方法による文節単位の構文解析を行なう際の重要な情報になる．

表7　$\pi_{1\sim(r-1)}(\sigma_{r>=c}$(表5))の AIC, χ^2, G^2 統計量

c	分布の数	比較の回数	同じであると認められた回数 AIC	χ^2	G^2
3	9	36	28	29	27
4	8	28	15	15	15
5	7	21	8	10	10
6	6	15	4	4	4
7	5	10	4	4	4
8	4	6	3	3	3
9	3	3	2	2	2
10	2	1	1	1	1
合計	44	120	55	68	66

表5に著者毎に分けられているものをレンジ単位に一つにまとめて用いた．

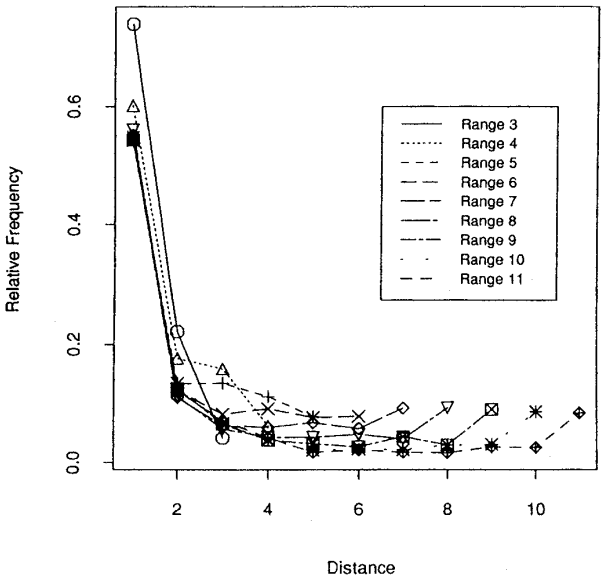

図3　レンジ毎の係り受け距離の分布

4．文節の係り受け距離と読点

本節では読点と文節の係り受け距離との関係について分析を行う．例えば次の文

私は，数人の作家の文章を，次に述べる方法で分析した．

の文節の係り受け距離は

距離1　数人の→作家
　　作家の→文章
　　次に→述べる
　　述べる→方法
　　方法で→分析した
　　分析した→．
距離4　文章を→分析した
距離7　私は→分析した

となる．多くの文の観察から，係り受けの距離4以上の成分の後ろに読点を打つと，標準的な打ち方になるという説もある．例としてあげた文は，係り受け距離が4またはそれ以

上の場合に読点を打ってある．もし，この規則に普遍性があるとすると，機械で生成される文，あるいは機械翻訳を行なう際，訳された日本語文章に機械的に読点を打つ方法の根拠を与えることが可能となる．

4.1 係り受け距離毎に読点を打つ比率

ある文章で（係り受け距離 j で読点を打った頻度）/（係り受け距離 j の頻度）を係り受け距離別に読点を打つ比率と呼ぶことにする．3人の21編の文章の距離別に読点を打つ比率を表8の下の行に示した．表8からわかるように，距離3～4以下で読点を打つ場合もあり，距離3～4以上でも読点を打たない場合もある．分析に用いた文章では，文節の係り受け距離が大きいとかならず読点を打つというような規則は見られない．また，同じの著者内でも距離毎に読点を打つ比率の分散が大きい．例えば，井上の距離3では読点を打つ比率が最も高いのは0.5310で，最も低いのは0.1544である．

表8 距離毎に読点を打ち頻度及び打つ比率

文章の記号	1	2	3	4	5	6	7	8	9	10
	3	18	30	34	38	31	13	18	18	18
I 1	0.002	0.064	0.170	0.256	0.409	0.508	0.464	0.720	0.783	0.667
	15	66	80	52	44	47	35	25	19	13
I 2	0.009	0.199	0.432	0.634	0.620	0.870	0.796	0.862	0.905	0.813
	16	42	57	54	54	36	27	25	21	15
I 3	0.010	0.138	0.368	0.482	0.628	0.610	0.750	0.735	0.913	0.790
	19	49	77	60	45	40	32	23	17	12
I 4	0.012	0.166	0.531	0.659	0.726	0.800	0.762	0.852	0.810	0.750
	4	15	21	31	16	23	13	17	8	9
I 5	0.003	0.087	0.154	0.337	0.302	0.535	0.542	0.680	0.381	0.818
	6	18	32	31	25	17	33	18	15	13
I 6	0.005	0.098	0.281	0.392	0.472	0.436	0.805	0.783	0.882	0.722
	5	16	29	25	25	16	21	10	19	10
I 7	0.004	0.068	0.242	0.317	0.472	0.500	0.724	0.476	0.679	0.833
	17	38	30	47	43	22	26	28	17	18
I 8	0.012	0.153	0.207	0.465	0.614	0.595	0.743	0.636	0.850	0.783
	6	19	29	39	29	28	23	26	11	7
M 1	0.003	0.057	0.169	0.315	0.382	0.549	0.511	0.867	0.407	0.539
	8	23	26	34	29	31	31	18	11	17
M 2	0.006	0.091	0.182	0.382	0.427	0.535	0.721	0.692	0.611	0.810
	7	32	37	44	41	17	19	10	15	12
M 3	0.005	0.107	0.236	0.400	0.578	0.370	0.594	0.556	0.625	0.632
	6	20	38	43	28	31	25	16	10	8
M 4	0.005	0.088	0.336	0.494	0.583	0.620	0.807	0.667	0.714	0.667
	12	32	42	38	35	30	28	14	11	7
N 1	0.011	0.158	0.342	0.521	0.556	0.811	0.824	0.737	0.917	1.000
	5	4	18	33	33	23	16	10	17	12
N 2	0.005	0.019	0.135	0.379	0.533	0.548	0.615	0.714	0.773	0.571
	4	18	17	32	25	27	20	15	15	10
N 3	0.003	0.066	0.114	0.320	0.424	0.519	0.625	0.577	0.652	0.833
	5	18	19	31	31	24	15	12	6	8
N 4	0.004	0.066	0.135	0.320	0.449	0.558	0.429	0.500	0.400	0.500
	8	19	37	19	37	18	21	7	15	3
N 5	0.006	0.084	0.250	0.232	0.569	0.546	0.568	0.412	0.682	0.500
	7	15	27	36	23	23	33	20	11	9
N 6	0.005	0.047	0.166	0.330	0.284	0.469	0.702	0.588	0.647	0.600
	10	24	35	28	24	24	21	22	12	17
N 7	0.006	0.100	0.212	0.286	0.414	0.453	0.467	0.733	0.632	0.708
	2	22	32	33	25	21	27	14	15	18
N 8	0.001	0.078	0.178	0.290	0.347	0.396	0.711	0.483	0.682	0.643
	1	18	22	22	27	17	15	19	7	6
N 9	0.001	0.065	0.122	0.220	0.403	0.340	0.484	0.633	0.467	0.375

4.2 距離毎に読点を打つ頻度と書き手の個性

読点の打ち方には明確な基準がないため，読点をどの文字の後に打つかに関しては書き手の個性が明確に現われるという結果が報告されている[(1),(2),(3)]．本節では係り受け距離毎に読点を打った頻度から考察を行うことにする．文章 i において距離 j で読点を打つ頻度（表8の上の行）を x_{ij} とすると，表8の距離毎に読点を打つ頻度のマトリックスは

$$X_{I\times J}=[x_{ij}]$$

で表記できる．距離毎に読点を打つ頻度が書き手によって異なるかを考察するため，まず $X_{I\times J}$ を用いて χ^2 距離（本稿の3.1節を参照）を求め，その群内，群間の距離を表9に示した．表9から，井上，中島の群内の距離が最小の群間の距離より大きいことがわかる．また，すべての χ^2 距離が $\chi^2(df=9,\ 5\%)$ の値14.68より大きく下まわる．これは，距離毎に読点を打つ頻度には書き手の個性がみられないことを意味する．

念のために，距離毎に読点を打つ頻度を著者毎にまとめ分析も行った．係り受け距離毎に打った読点の頻度・相対頻度を著者毎にまとめ，表10に示した．表10のデータにおける AIC，χ^2，G^2 統計量はそれぞれ13.48，22.54，22.52で，3人の距離毎に読点を打つ頻度には明確な差があると統計的に認められない．

係り受け距離毎に読点を打つ頻度の分布に書き手の個性が見られないことと筆者らが発表した論文[(1),(2),(3)]の結果とを合わせて考えると非常におもしろい．もっと深く追求する価値があると思う．

表9 係り受け距離毎に読点を打つ頻度のデータを用いた場合の χ^2 距離の平均値

著者	群内	群		間	最小の群間
井上	7.169	0.000	6.035	7.197	6.025
三島	5.706	6.035	0.000	5.845	5.845
中島	7.365	7.197	5.845	0.000	5.845

5．まとめ

本稿では，文節の係り受け関係を数値化した係り受け距離の分布について統計分析を行い，以下のようなことを明らかにした．

- 係り受け距離の分布には書き手の個性が明確には現われない．
- 係り受け距離の分布はレンジ毎に分けていない場合は，書き手及び文章に関係なく同じL形分布（あるいはZipf法則）に従う．しかし，レンジ毎に分けた場合は，書き手とは無関係であるが，レンジ毎の分布は異なるため，Zipf法則あるいはL型分布でモデル化するには無理がある．
- レンジの値が大きくなると（約6以上）最後の文節に係る確率が目立つ．このような現象はレンジの値が大きくなるほど明確である．例えば，レンジの値が10の場合，その文節が最後の文節に係る確率は直後の第3文節に係る確率より大きい．
- 必ずしも係り受け距離が遠くなるとすべて読点を打つに限らない．
- 係り受け距離毎に読点を打つ頻度には書き手の個性が見られない．

本研究の継続としては，より多くの文章，

表10 係り受け距離と読点を打つ頻度

距離	1	2	3	4	5	6	7	8	9	10
井上	86	262	356	334	290	232	200	164	134	108
三島	27	94	130	160	127	107	98	70	47	44
中島	60	190	287	315	288	238	221	149	119	98

特に規範的な文（例えば，小中学校のテキスト等）について分析をさらに進めるのは興味深い．また，係り受け距離と読点の打ち方に関しては単に文節の係り受け距離だけではなく，文節の成分（品詞など）との関係なども考慮して分析を進める必要があると思う．このような点については今後の研究課題として続報としたい．

謝辞 本研究は神戸学院大学の樺島忠夫教授のご指導がなかったら始まることがなかったと思います．本研究に用いた文節の係り受け関係に関する具体的な規則は樺島教授が作成されたものを使わせていただきました．計量分析に用いたデータベースは文部省統計数理研究所及び総合研究大学院大学の村上征勝教授の研究費で作成しました．心より感謝いたします．

注

（注１）コンピュータで文のなかから情報を抽出する際，文節の番号についてはプログラムによりカウントすることが可能であるため，データベースには文節番号を記入しなくてもよい．

（注２）本稿で用いているレンジの定義は，統計学でデータのバラツキを評価するために用いるレンジ（range）とは意味が異なる．

参考文献

(1) 金 明哲，村上征勝 (Jin, M.Z and Murakami, M.): *'Authors' Characterristic Writing Styles as Seen Through Their Use of Commas'*, Behaviormetrika, Vol.20, 63-76, (1993a).

(2) 金 明哲，樺島忠夫，村上征勝：読点と書き手の個性，『計量国語学』，Vol.18，382-391，(1993 b)．

(3) 金 明哲：読点の打ち方と文章の分類，『計量国語学』，Vol.19，No.7，317-330，(1994)．

(4) 黒橋貞夫，長尾 真：並列構造の検出に基づく長い日本語の構文解析，『自然言語処理』，Vol.1，No.1，pp.35-58，(1994)．

(5) 坂元・石黒・北川：『情報量統計学』，共立出版株式会社，pp.65-77（1991）．

(6) 武石英二，林 良彦：接続構造解析に基づく日本語複文の分割，『情報処理論文誌』，Vol.33，No.5，pp.652-663，(1992)．

(7) 任 福継・他：日中機械翻訳における係り受け構造の可保留曖昧関係について，『情報処理』，Vol.34，No.8，(1993)．

(8) Maruyama, H. and Ogino, S.: *'A Statistical Prorerty of Japanes Phrareto-Phrase Modifications'*, Mathematical Linguistics, Vol.18, No.7, pp.348-352, (1993).

(9) 横田一正・宮崎収兄：『新データベース論』，共立出版株式会社，pp.39-50，(1994)．

(10) 吉田将：二文節間の係り受けを基礎とした日本語文の構文解析，『電子通信学会論文誌』，Vol.55-D, No.4, pp.238-244(1971)．

1995 年 12 月 26 日受付
1996 年 2 月 13 日受理