2007－CH－75 (8)
2007／7／27

中古散文22作品の類似度の測定

前川 武
大阪国際大学短期大学部 国際文化学科

日本語の文献に関する研究においては、本文テキストのデータベース化と語彙索引の電子化が進めばさまざまな角度からの計量的分析が可能になる。
古典文学の分野では、本文テキストのデータベース化は進んでいるが、索引の電子化は、なかなか進んでいないのが現状である。このような現状の中で、村田菜穂子氏は、古代語の形容詞と形容動詞について、単なる作品における語の出現度数、用例だけではなく、詳細な情報を付加した語彙表を作成し、様々な計量分析を行っている。
今回、村田氏の作成したデータに基づき、中古の散文資料 22 作品について、その語彙の使用状況から作品間の類似度を測定する試みを行った。

A trial to measure a similar degree of 22 Old and Medieval prose works.

Takeshi Maekawa
Osaka International College, Department of Intercultural Relations.

In the research on a Japanese document, if the computerization of a full text database and glossarial index advances, measurement analysis from various angles becomes possible. The computerization of the glossarial index is not advancing easily, though that of a full text database is advanced.
In such a situation, Nahoko Murata makes the glossarial index of the adjective and the adjective verb of an ancient word, and is doing various analyses and consideration.
I tried to measure a similar degree of 22 Old and Medieval prose works by using the data in the writing of Murata.

I はじめに

日本語の文献に関する研究においては、本文テキストのデータベース化と作品中に使用されている語の属性、出現頻度などをまとめた語彙索引の電子化が進めばさまざまな角度からの計量的分析が可能になる。
日本語の古典文学の分野では、国文学研究資料館が、利用目的を研究の範囲に限定し、さらに登録制にした上で、岩波書店旧版『日本古典文学大系』全作品（100巻560作品）の全文データベースとその検索システムを試験的に公開している。これを利用すると、複数の作品を対象とした文字列検索や文字頻度分析などが簡単にできる。[1)]
このように、本文テキストのデータベース化は進んでいるが、それに比べて、索引の電子化は、なかなか進んでいないのが現状である。その理由の一つとして、印刷物の索引をベースにする場合、複合語の認定基準が索引作成者の判断に左右されるところが大きいことがあげられる。そのほかにも、従来の索引にない語の属性を付加したい場合など、再度本文を調査する必要があることなどがあげられる。
このような現状の中で、村田菜穂子氏は、古代語の形容詞と形容動詞について、単なる作品における語の出現度数、用例だけではなく、語構造・単位数・結合タイプ・結合型といった語構造

に関わる属性、造語形式・造語型といった造語論に関わる属性、活用の種類などの詳細な情報を付加した語彙表を作成し、古代語形容詞・形容動詞の活用別、出現時代別、結合タイプ別あるいはこれらを複合した形での延べ語数・異なり語数の分布状況を分析し、それらの史的変遷について考察を行い、『形容詞・形容動詞の語彙論的研究』[2]の中で分析・考察内容を明らかにしている。

今回、同書に別表として掲載されているデータに基づき、中古の散文資料22作品[3]について、その語彙の使用状況から作品間の類似度を測定する試みを行った。

Ⅱ　類似度の測定の方法

語彙の使用状況から作品間の類似度を測定する方法については、これまでいくつかの手法が考案されてきた。

宮島達夫氏は、「語いの類似度」[4]において、見出し語数の共通度に注目した水谷静夫氏の式[5]と安本美典・本多正久氏の式[6]について解説しながら、独自の説を展開し、様々な分野への適用を試みている。

以下に、宮島氏の手法を簡単に説明する。

見出し語 $W_1,W_2,W_3,\cdots,W_n$ があったとき、見出し語 W_i の作品Aにおける使用率を $P_i(A)$、見出し語 W_i の作品Bにおける使用率を $P_i(B)$で表し、両者の小さい方を $\min(P_i(A),P_i(B))$で表すと、以下の式で算出される値 C_{AB} が作品AB間の類似度となるというものである。

$$C_{AB}=\sum_i \min(P_i(A),P_i(B))$$

この式は、不一致度の合計を用いた次の式から導き出されている。

$$C_{AB}=1-\frac{1}{2}\left(\sum_i \left(\left|P_i(A)-P_i(B)\right|\right)\right)$$

さらに、宮島氏は、この類似度を基にして、複数作品間の近さを図示する方法を考案している。それは、異なる2作品間の不一致度（1－類似度）をすべて算出し、その中で一番高い値をとったときの2作品の組合せ（例えばAとB）を求め、その不一致度を底辺の長さとする三角形を想定し、その両端に2つの作品を配置し、比較したい作品（例えばCとD）との不一致度を算出し、それぞれの長さを残りの2辺とし、その頂点に比較したい作品を配置するというものである。これを図示すると図1のようになる。

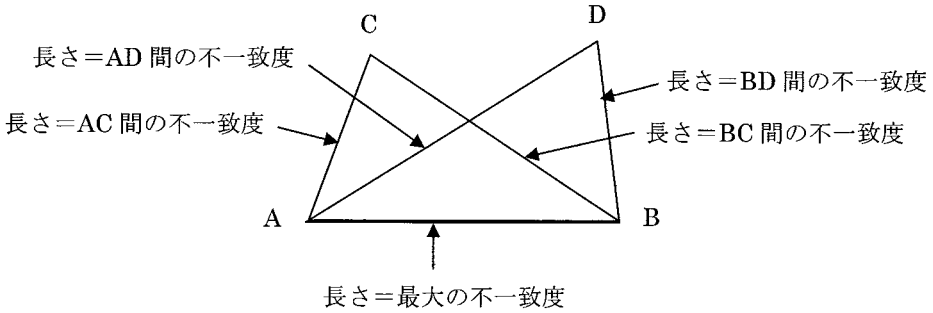

図1　複数作品間の近さを図示する方法

今回の類似度の測定にあたっては、この宮島氏の式および図示の方法を使うこととした。

Ⅲ　類似度の測定

1　対象となるデータ

今回対象としたデータは、『形容詞・形容動詞の語彙論的研究』[2)] の別表三「中古散文作品の形容詞対照語彙表」収録の形容詞 1124 語および別表四「中古散文作品の形容動詞対照語彙表」形容動詞 1093 語とした。

データの内容は、見出し語と中古散文 22 作品での述べ語数（出現回数）である。参考までに形容詞のデータの例を表 1 に示す。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	見出し語	漢字	活用	竹取	土佐	伊勢	平中	大和	多武峯少		宇津保	蜻蛉	落窪	和泉	枕	源氏	紫式部	堤	寝覚	浜松	更級	狭衣	大鏡	讃岐	とりか	総計
2	あいぎょうな	愛敬無	ク										5		2	5						1	1			14
3	あいなし	(空白)	ク						1		10	8	10	2	9	101	4	4	34	14	1	1			17	216
4	あうなし	奥無	ク													8	2		3			1				14
5	あかし	赤	ク				1				18	1			15	13		1	4	4		8	5		1	71
6	あかし	明	ク		2			3	3		13	7	4	9	18	25	2	3	16	11	15	12	7	2	6	159
7	あかしがた	明難	ク											1		2						2				5
8	あきたし	飽・厭	ク													2			2			1				5
9	あきらけし	明	ク													1							1			2
10	あきれいた	呆痛	ク																2							2
11	あくがれが	憧難	ク													1										1
12	あけがたし	開難	ク								2			1												3
13	あけやすし	明易	ク								1															1
14	あさし	浅	ク		1			4			11			2	2	94	1	3	49	46		13	1		30	257
15	あさまし	浅	シク	2		1	2	7	9			33	19	11	29	200	7	14	71	55	3	109	34	7	77	690
16	あざれがま	戯	シク													2										2
17	あざればま	戯	シク														1									1
18	あし	悪	シク	4	8	4	2	6		1	80	31	44	5	41	95	7	3	29	11	2	28	29	6	12	448
19	あそびよし	遊良	ク																	1						1
20	あだあだし	徒徒	シク								1			2		4	1			1		2			1	12
21	あだし	徒	シク													1										1
22	あたらし	新	シク						1		19	2	4		12	17	3	2	2	1	1	2	2		2	70
23	あたらし	惜	シク									1	3			31			8	7		3	3		4	60
24	あぢきなし	味気無	ク			1					35	4	1	2	8	66		3	17	2		24		3	2	168
25	あぢなし	味無	ク																						1	1
26	あつかはし	(空白)	シク								1				1	5						1				8
27	あつし	厚・敦	ク								3				4	2						1	1			11
28	あつし	暑・熱	ク			2					17	5			14	16		1	4	2	1	9	6	2	4	83
29	あつし	篤	シク								5					8							1		1	15
																～途中省略～										
1120	ぎりうし	折憂	ク					1			1															2
1121	ぎりにくし	折難	ク										1													1
1122	ぎりよし	折良	ク													3			1							4
1123	ぎをし	謀謀	シク												2	11									4	17
1124	ぎんなし	女	シク													7									1	8
1125	ぎんなどほし	女遠	ク													1										1
1126				278	179	304	319	685	149	91	9027	1761	2298	438	3525	22316	728	750	5807	3462	585	5104	1640	458	3320	63224

表 1　測定対象となるデータの例（中古散文作品の形容詞 1124 語）

2　各作品の使用率の算出

データの件数が多いため、Excel VBA を使って算出した。

3　2 作品間の類似度の算出

22 作品の異なる 2 作品の組合せは ${}_{22}C_2=231$ とおりと多いので、これも Excel VBA を使って算出した。

4　2 作品間の類似度および不一致度のまとめ

3 で求めた値を見やすい形に整理するとともに、類似度の最小値、類似度が最小になる 2 作品の組を求め、さらに、その 2 作品とそれらを除いた 20 作品との類似度および不一致度を求めるもので、これも Excel VBA を使って算出した。

5　22 作品の距離感の図示

不一致度の最大の組合せの 2 作品とその 2 作品と他の 20 作品との不一致度が求まったので、あとは、20 作品のそれぞれについて、（最大の不一致度，不一致度最大の組合せの一方との不一致度，不一致度最大の組合せのもう一方との不一致度）を 3 辺とする三角形の頂点の位置を求めて点をプロットしていけばよい。

しかし、Excel の標準のグラフ作成機能には、このような場合に適合するグラフがなく、変数を変換して適用するにもやはり適切なものがない。

また、Excel VBA のユーザーフォームでは、グラフィックス系のコントロールが使用できない。

別アプリケーションや DLL を呼び出すこともできなくはないが、環境に依存することになるため、できれば使用したくない。

シートの倍率を縮小して、セルのサイズを小さくして、ドットの代わりにする方法も考えたが、これだと文字を表示したときに小さすぎて見えなくなってしまう。

そこで、図形のオートシェープの楕円を Excel VBA からコントロールして、楕円をドットの代わりに使って 22 作品の点をプロットしていくこととした。

Ⅳ　結果および考察

1　形容詞の使用率から見た中古散文 22 作品の類似度

まず、異なる 2 作品間での類似度および不一致度の算出結果は、表 2 のようになった。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
2	0.522																				
3	0.496	0.492																			
4	0.494	0.476	0.512																		
5	0.497	0.486	0.554	0.584																	
6	0.366	0.322	0.369	0.440	0.484																
7	0.458	0.403	0.470	0.436	0.493	0.395															
8	0.557	0.486	0.529	0.523	0.646	0.401	0.512														
9	0.507	0.479	0.513	0.550	0.603	0.435	0.421	0.610													
10	0.517	0.457	0.440	0.467	0.570	0.380	0.409	0.656	0.592												
11	0.410	0.384	0.421	0.525	0.486	0.388	0.415	0.486	0.580	0.499											
12	0.442	0.395	0.423	0.450	0.505	0.283	0.355	0.588	0.516	0.603	0.429										
13	0.453	0.370	0.440	0.465	0.552	0.361	0.363	0.637	0.615	0.632	0.540	0.529									
14	0.445	0.383	0.452	0.461	0.502	0.298	0.360	0.589	0.537	0.562	0.486	0.591	0.593								
15	0.489	0.406	0.451	0.529	0.578	0.390	0.390	0.614	0.614	0.646	0.555	0.605	0.650	0.566							
16	0.499	0.409	0.467	0.464	0.504	0.370	0.359	0.557	0.592	0.584	0.519	0.500	0.674	0.520	0.613						
17	0.516	0.402	0.438	0.433	0.516	0.389	0.395	0.591	0.566	0.580	0.486	0.521	0.638	0.509	0.601	0.707					
18	0.459	0.416	0.439	0.454	0.514	0.372	0.356	0.528	0.542	0.506	0.489	0.538	0.494	0.510	0.563	0.501	0.572				
19	0.500	0.407	0.447	0.486	0.546	0.430	0.400	0.643	0.645	0.599	0.562	0.531	0.710	0.542	0.643	0.684	0.687	0.525			
20	0.514	0.406	0.418	0.426	0.560	0.360	0.431	0.641	0.541	0.599	0.430	0.630	0.562	0.511	0.588	0.530	0.562	0.534	0.584		
21	0.545	0.450	0.487	0.482	0.507	0.394	0.392	0.568	0.541	0.517	0.465	0.510	0.489	0.476	0.535	0.536	0.538	0.505	0.602	0.529	
22	0.481	0.385	0.414	0.448	0.518	0.414	0.390	0.597	0.593	0.620	0.544	0.499	0.689	0.508	0.638	0.703	0.718	0.530	0.711	0.548	0.531

	高い	低い			1	2	3	4	5	7	8	9	10	11	13	14	15	16	17	18	19	20	21	22
1	0.718	0.283		6	0.366	0.322	0.369	0.44	0.484	0.395	0.401	0.435	0.38	0.388	0.361	0.298	0.39	0.37	0.389	0.372	0.43	0.36	0.394	0.414
2	0.711	0.298			0.634	0.678	0.631	0.56	0.516	0.605	0.599	0.565	0.62	0.612	0.639	0.702	0.61	0.63	0.611	0.628	0.57	0.64	0.606	0.586
3	0.710	0.322																						
4	0.707	0.355		12	0.442	0.395	0.423	0.45	0.505	0.355	0.588	0.516	0.603	0.429	0.529	0.591	0.605	0.5	0.521	0.538	0.531	0.63	0.51	0.499
5	0.703	0.356			0.558	0.605	0.577	0.55	0.495	0.645	0.412	0.484	0.397	0.571	0.471	0.409	0.395	0.5	0.479	0.462	0.469	0.37	0.49	0.501
				0.283																				
				0.717																				

表 2　中古散文 22 作品の形容詞の使用率から見た類似度および不一致度

上部の階段状のものは、横軸の 1～21 が 2 作品の組合せの一方を表し、縦軸の 2～22 がもう一方を表し、その交点に対応する数値が 2 作品間の類似度を表す。

例えば、横軸が 3、縦軸が 5 の交点を見ると値は 0.554 であるが、これは、作品 3 と作品 5 の類似度が 0.554 であることを表す。なお、類似度の平均は 0.506、標準偏差は 0.087 であった。

下部の左側の 2 列は、全体の中で最も高い類似度と最も低い類似度を 5 つずつ示している。

下部の右側部分の内、最左列は、全体の中で最も類似度の低かったものが作品 6 と作品 12 の組合せであったこと、そのときの類似度が 0.283、不一致度が 0.717 であることを示している。

下部の右側部分の内、最上部の行は、22 作品から最も類似度の低かった組合せの作品 6 と作品 12 を除いた 20 作品を示している。

その下の 2 行は、作品 6 と上記 20 作品との間の類似度（上段）と不一致度（下段）を示す。

さらに 1 行空けたあとの 2 行は、作品 12 と上記 20 作品との間の類似度（上段）と不一致度（下段）を示す。

表 2 では、作品を作品名ではなく 1～22 の連番で表しているが、これは、成立年代の順を反映しているもので、年代の推移を念頭に置いた場合、この方がわかりやすいこともあって、敢えて連番で表記した。

参考までに中古散文 22 作品の作品名と連番との関係を表 3 に示す。

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
竹取物語	土佐日記	伊勢物語	平中物語	大和物語	多武峯少将物語	篁物語	宇津保物語	蜻蛉日記	落窪物語	和泉式部日記	枕草子	源氏物語	紫式部日記	堤中納言物語	夜の寝覚	浜松中納言物語	更級日記	狭衣物語	大鏡	讃岐典侍日記	とりかへばや物語

表3　中古散文22作品の作品名と連番との関係

表2の中で類似度の高いものは、次の5つである。

17－22：0.718　　19－22：0.711　　13－19：0.710　　16－17：0.707　　16－22：0.703

逆に低いものは、次の5つである。

6－12：0.283　　6－14：0.298　　2－6：0.322　　7－12：0.355　　7－18：0.356

次に、22作品間の距離感は図2のような結果になった。

図3は、図2の上部を拡大したものである。

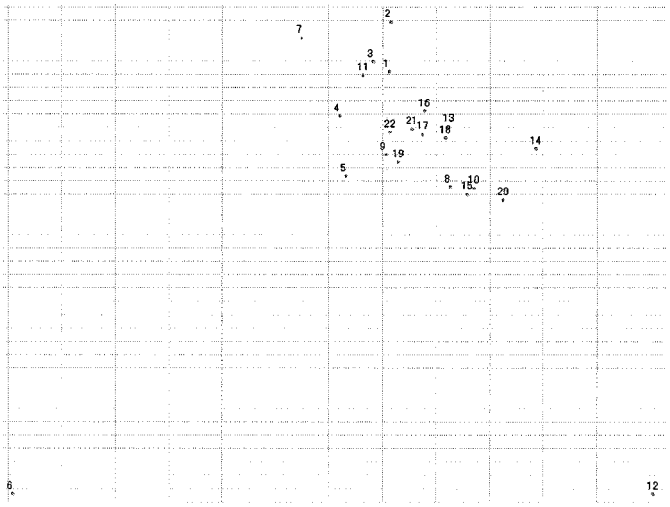

図2　中古散文22作品の形容詞の使用率から見た距離感

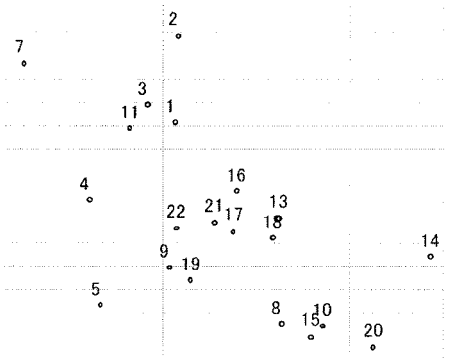

図3　図2の上部を拡大したもの

近いものを集めてみると、次のようになる。

グループ 1：8,10,15

グループ 2：9,13,16,17,18,19,21,22

グループ 3：1,3,11

しかし、これらのグループに属する 2 作品間の類似度を表 2 で確認すると、図 2 の位置関係と符号しないものが出てくる。

例えば、17－22 の類似度が 0.718、17－21 の類似度が 0.538 であることからわかるように、数値的には 17 から見れば 22 の方が近いにも関わらず、図 2 では 22 の方が遠くなっている。あるいは、8－10 の類似度が 0.656、10－15 の類似度が 0.646 であるのに対して、図 2 上では、8－10 の距離＞10－15 の距離となっているなどである。

つまり、図 2 は、あくまでも底辺にあたる 6 および 12 と他の頂点との三者間の距離を表しているだけで、ある頂点と別の頂点との距離を間接的に表しているわけではないことがわかる。

次に、表 2 の値に基づき、三者が相互に高い類似度を示すものを一つのグループとして位置づけてみると、次のようになる。（ ）内が作品番号の組合せ、[ ] 内が類似度の範囲である。

（13,19,22）[0.689－0.711]、（17,19,22）[0.687－0.718]、（13,16,19）[0.674－0.710]

（16,17,22）[0.703－0.718]、（10,13,15）[0.632－0.650]、（8,10,13）[0.637－0.656]

これを見ると、13,16,17,19,22 は相互に類似度が高い関係にあることがわかり、上のグループ 2 とかなりの部分で重なることも見てとれる。

2　形容動詞の使用率から見た中古散文 22 作品の類似度

まず、異なる 2 作品間での類似度および不一致度の算出結果は、表 4 のようになった。

2	0.226																					
3	0.273	0.255																				
4	0.321	0.324	0.393																			
5	0.283	0.294	0.393	0.485																		
6	0.164	0.189	0.219	0.305	0.405																	
7	0.136	0.116	0.238	0.301	0.406	0.392																
8	0.318	0.225	0.347	0.395	0.529	0.393	0.325															
9	0.179	0.247	0.291	0.352	0.495	0.363	0.39	0.524														
10	0.238	0.174	0.242	0.288	0.495	0.357	0.427	0.53	0.683													
11	0.225	0.219	0.27	0.38	0.466	0.432	0.34	0.458	0.5	0.441												
12	0.241	0.231	0.275	0.332	0.464	0.307	0.318	0.525	0.478	0.533	0.428											
13	0.249	0.201	0.319	0.287	0.388	0.295	0.269	0.566	0.526	0.506	0.431	0.517										
14	0.197	0.196	0.305	0.273	0.335	0.24	0.194	0.401	0.377	0.385	0.311	0.432	0.493									
15	0.238	0.185	0.268	0.309	0.46	0.322	0.31	0.506	0.467	0.544	0.455	0.524	0.523	0.413								
16	0.252	0.198	0.272	0.271	0.39	0.291	0.264	0.499	0.497	0.509	0.433	0.483	0.663	0.437	0.508							
17	0.288	0.25	0.329	0.356	0.472	0.373	0.318	0.555	0.511	0.499	0.505	0.543	0.641	0.406	0.541	0.642						
18	0.224	0.202	0.315	0.316	0.434	0.357	0.307	0.45	0.452	0.425	0.487	0.446	0.458	0.361	0.446	0.471	0.547					
19	0.25	0.24	0.285	0.334	0.42	0.342	0.332	0.536	0.52	0.528	0.457	0.548	0.623	0.39	0.551	0.61	0.626	0.415				
20	0.279	0.305	0.343	0.321	0.356	0.248	0.255	0.435	0.423	0.401	0.376	0.464	0.464	0.379	0.397	0.437	0.464	0.37	0.479			
21	0.177	0.191	0.177	0.222	0.315	0.274	0.258	0.383	0.365	0.385	0.376	0.409	0.404	0.277	0.397	0.416	0.438	0.364	0.457	0.382		
22	0.262	0.19	0.278	0.273	0.42	0.326	0.312	0.514	0.511	0.625	0.476	0.499	0.637	0.427	0.537	0.669	0.655	0.461	0.594	0.397	0.415	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	

	高い	低い			1	3	4	5	6	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0.669	0.116		2	0.226	0.255	0.324	0.294	0.189	0.225	0.247	0.174	0.219	0.231	0.201	0.196	0.185	0.199	0.25	0.202	0.24	0.305	0.191	0.19
2	0.663	0.136			0.774	0.745	0.676	0.706	0.811	0.775	0.753	0.826	0.781	0.769	0.799	0.804	0.815	0.802	0.75	0.798	0.76	0.695	0.809	0.81
3	0.655	0.164																						
4	0.642	0.174		7	0.136	0.238	0.301	0.406	0.392	0.325	0.39	0.427	0.34	0.318	0.269	0.194	0.31	0.264	0.318	0.307	0.332	0.255	0.258	0.312
5	0.641	0.177			0.864	0.762	0.699	0.594	0.608	0.675	0.61	0.573	0.66	0.682	0.731	0.806	0.69	0.735	0.682	0.693	0.668	0.745	0.742	0.688
				0.116																				
				0.884																				

表 4　中古散文 22 作品の形容動詞の使用率から見た類似度および不一致度

類似度の平均は 0.381、標準偏差は 0.119 であった。

表 4 の中で類似度の高いものは、次の 5 つである。

15－22：0.669　　13－16：0.663　　17－22：0.655　　16－17：0.642　　13－17：0.641

逆に低いものは、次の 5 つである。

2－7：0.116　　1－7：0.136　　1－6：0.164　　2－10：0.174　　1－21：0.177

次に、22 作品間の距離感は図 4 のような結果になった。

図 5 は、図 4 の上部を拡大したものである。

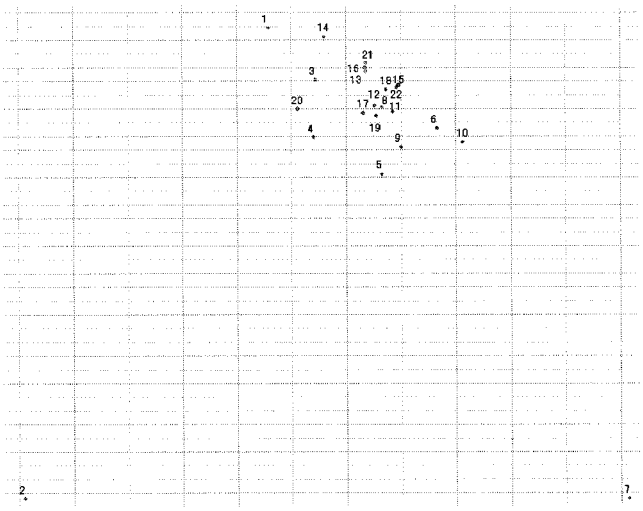

図 4　中古散文 22 作品の形容動詞の使用率から見た距離感

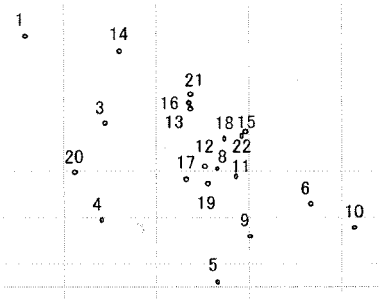

図 5　図 4 の上部を拡大したもの

近いものを集めてみると、次のようになるが、形容詞の場合と同様の問題がある。

グループ 1：5,6,9,10

グループ 2：11,13,15,16,17,18,19,21,22

グループ 3：3,4,14,20

次に、表 4 の値に基づき、三者が相互に高い類似度を示すものを一つのグループとして位置づけてみると、次のようになる。（　）内が作品番号の組合せ、[　] 内が類似度の範囲である。

(16,17,22) [0.642－0.669]、(13,17,22) [0.637－0.655]、(13,16,22) [0.637－0.669]

これを見ると、13,16,17,22 は相互に類似度が高い関係にあることがわかり、上のグループ 2 とかなりの部分で重なることも見てとれる。

3　形容詞の使用率から見た場合と形容動詞の使用率から見た場合との比較

図 2、3 と図 4、5 を見比べると、多くの部分で共通点が見られる。

例えば、6 と 7 は、年代的に近いにも関わらず、形容詞、形容動詞とも距離が大きく隔たっていること、13,16,17,22 が極めて近い位置にあることなどがあげられる。

しかし、ここでも、1 で述べた問題点があることを忘れてはいけない。

V　おわりに

今回は、宮島氏の提唱した手法に基づいて、形容詞および形容動詞の見出し語の使用率から中古散文22作品の距離感を測定して図示するプロセスをExcel VBAと使って実現してみた。

その結果、一つの平面上に図示することはできたが、距離感に矛盾が生じることも明らかになった。今後は、他の分析手法との比較などを通して、これらの矛盾を解消していきたい。

また、今回は、見出し語の使用率に着目したが、形容詞や形容動詞の個体そのものの持つ属性と用いられている作品との関係、作品の特性や作者の特性などについても着目し、分析を行っていきたい。

注

1)国文学研究資料館　http://www.nijl.ac.jp/index.html

2)村田菜穂子、『形容詞・形容動詞の語彙論的研究』、初版、和泉書院、2005年

3)『竹取物語』、『土佐日記』、『伊勢物語』、『平中物語』、『大和物語』、『多武峯少将物語』、『篁物語』、『宇津保物語』、『蜻蛉日記』、『落窪物語』、『和泉式部日記』、『枕草子』、『源氏物語』、『紫式部日記』、『堤中納言物語』、『夜の寝覚』、『浜松中納言物語』、『更級日記』、『狭衣物語』、『大鏡』、『讃岐典侍日記』、『とりかへばや物語』

4)宮島達夫、「語いの類似度」、『國語學』、82、1970年

5)2作品の見出し語の数をa,b、共通の見出し語の数をxとしたとき、類似度 $=\dfrac{x}{a+b-x}$

6)2作品の見出し語の数をa,b、共通の見出し語の数をxとしたとき、類似度 $=\dfrac{x}{\sqrt{ab}}$

参考文献

1)宮島達夫、「総索引への注文」、『國語學』、76、1969年

2)村上征勝・金明哲、『講座　人文科学研究のための情報処理〔第5巻　数量的分析編〕』、初版、尚学社、1998年

3)C&R研究所、『超図解 EXCEL VBAハンドブック EXCEL 2000/2002/2003対応』、初版、エクスメディア、2004年