テキストマイニングによる レポート課題の分析

—— 文章と構成の観点から ——

植　田　　麦

はじめに

教員として授業を担当していると、「今年は教育効果が高い」と感じられることがある。しかし、受講生全体の意欲や能力が高かったのか、たまさか授業参加に積極的な学生が数名いたために稿者の認知バイアスが働いたのか、あるいは他の要因による錯覚なのか、いずれにせよほぼすべてのケースにおいて実証が難しい。これはなにも、稿者の担当しているような「国語表現」、いわゆる文章表現法（アカデミックライティング）の授業に限ったことではないだろう。

また、一定期間経過時点に課したレポートをみたとき、授業開始時点に課したものと比較して全体に表現能力や構成力の向上が感じられるとして、それがどの程度の向上であったのかを知ることも難しい。さらに、授業担当者として受講生の能力をより向上させられたのではないか、と反省することもあるが、それがいかなるもので、かつ、いかほどであったのかと問うてもその答えを得られることはほとんどない。

こうした、経験に伴う感覚と事実との懸隔は、教員として容易に埋めがたい苦悩のひとつであろう。

さて、稿者は先稿（植田麦：2020）において、2020 年度に担当した「国語表現」の教育効果測定についての研究成果を示した。上に教員としての苦悩について述べたが、当該研究は稿者なりにそのような苦悩を解決する一助としたつもりである。

しかし､そこではいくつかの知見を得られたと同時に､課題も浮き彫りとなった。そのため本稿は先稿を承けて、2020 年度春学期開講「国語表現」および 2021 年度春学期開講「国語表現」において受講生に課したレポートをもとに、再度その教育効果の測定を試みるものである。

1. 先稿の確認

論考に先立ち、先稿で用いた技術の確認を行い、得られた結論についてみておく。

分析では、テキストマイニングツールである MTmineR および KH Coder を用いた。また、これらのツールを共通して利用するために、独自にレポート用フォーマットを作成した[補注]。

先稿では､「国語表現」の受講により受講生は､

1. 語彙が豊富になる
2. 文章の体裁が整う
3. 文章自体の作成能力が向上する

の 3 点について変化がみられると仮定した。そのため、

1. 語彙の豊富さの計測と観測
2. 段落冒頭の字下げ・常体の使用・段落数・文の数等の計測と観測
3. 1 文あたりの文字数・高使用頻度語の計測と観測

を行い、その仮説の妥当性を検証した。

第一の仮説「語彙が豊富になる」については、必ずしもそうとはいえない結果がみられた。ただし、これについては他の年度との比較を行うことで、一回性の現象であるのかあるいは常態的であるのかを観測するべきであると結論づけた。

第二の仮説「文章の体裁が整う」については､教育効果をみることができた。また、指導上の有益性も確認できた（そのため、2021 年度はより積極的に指導内容に組み込んだ)。

第三の仮説「文章自体の作成能力が向上する」については、十分な教育効果をあげていないことが明らかとなった（これについても、2021 年度は指導上の課題として授業設定を行った)。

以上の確認に基づき、2020 年度「国語表現」と 2021 年度「国語表現」とで課したレポート課題を対照し、分析研究を行う。

2. 本稿における研究課題

研究対象とする授業内容とレポート課題の概要を示す。2020 年度は授業開始時期の遅れもあり、授業回数自体が予定よりも少なくなった。一方、2021 年度は 2700 分が確保された。ただし、授業の構成としては、大きな変更を行っていない。以下に示す第 1 グループから指導を始め、第 4 グループで指導を終えている。

第 1 グループ：同音異義語やアカデミックワードなど、主として「語」の理解に関わる内容

第 2 グループ：文のねじれや多義文など、主として「文」の作成に関わる内容

第 3 グループ：資料の読解やレポートの論理構成など、主として「構成」に関わる内容

第 4 グループ：学習内容に基づいた総括的内容

2020 年度は学期中に 6 回のレポート提出を課した。2021 年度も同様である。レポートを課す時期についても、ほぼ同様である。ただし、第 1 グループの初期と第 3 グループの中期に課すレポートのテーマを入れ替えた。すなわち、

2020A：2020 年度第 1 グループ初期 「日本において、一定年齢以上の独身者に対する税（いわゆる「独身税」）を課すべきか」字数：800 字以上 1600 字以内（レポート総数 87 件）

2020B：2020 年度第 3 グループ中期 「選択的夫婦別姓を認めるべきか」字数：800 字以上 1600 字以内（レポート総数 84 件）

2021A：2021 年度第 1 グループ初期 「選択的夫婦別姓を認めるべきか」字数：800 字以上 1600 字以内（レポート総数 84 件）

2021B：2021 年度第 3 グループ中期 「日本において、一定年齢以上の独身者に対する税（いわゆる「独身税」）を課すべきか」字数：800 字以上 1600 字以内（レポート総数 81 件）

とした。つまり、2020A と 2021B、2020B と 2021A とがそれぞれ同一のテーマ設定である。これにより、授業の進展によって生じるレポートの構成の変化を観測することが可能となると考えた。

以下の論考では、上の 2020A・2020B・2021A・2021B のレポートを比較し、先稿の課題を再検討する。すなわち、

1. 語彙が豊富になる
2. 文章の体裁が整う

3. 文章自体の作成能力が向上する

である。さらに、2020 年度「国語表現」と 2021 年度「国語表現」との比較により、

4. 適切な論理構成が可能となる

を新たな課題として設定したい。

3. 先稿との比較

3.1 語彙の豊富さ

以下に、4 回のレポートにおける異なり語数と延べ語数、語彙の豊富さの指標としての s の数値を示す。

表 1 語彙の豊富さ

	異なり語数	延べ語数	s
2020A	2693	20547	0.843
2020B	2412	21943	0.830
2021A	2593	19489	0.843
2021B	2410	20676	0.833

なお、ここで 2020 年度に課したレポートについて先稿と数字が異なるのは、KH coder のバージョンアップ・形態素解析器の変更・強制抽出語の設定・否定表現の抽出・表記揺れの吸収等を行ったことによる。ここでいう否定表現の抽出とは、たとえば「わから（ない）」を「わかる（否定）」のように、「わかる」とは別の語として抽出することである。否定表現の抽出については、KH Coder のプラグイン「否定表現チェッカー」を使用した。また、表記揺れの吸収についても、KH Coder のプラグインである「表記ゆれ＆同義語エディター」を使用した。

s については先稿に示したが、本稿でも簡略に述べる。語彙の豊富さを計測する指標としては異なり語数を延べ語数で割った TTR（Type Token Ratio）が広く知られる。しかしながら、文章量が増えるほど延べ語数が増加するのに対し、異なり語数は延べ語数と同じ比率では増加しない。そのため、TTR は文章量の異なるテキストを比較する指標としては不適切である。

この TTR を補正するために様々な指標が考案された。鄭弯弯・金明哲（2018）は複数の指標を比較検討し、s が最も文章量増減の影響をうけにくいと述べている。そのため、先稿と同様、本稿でも語彙の豊富さをはかる指標として s を

用いる。

2020 年度と 2021 年度のレポートにおける s をみると、変化現象としては一致している。先稿でも確認したとおり、2020 年度は 2020A から 2020B にかけて s の低下がみられる。また、2021 年度も同様で、2021A から 2021B にかけて s は低下している。数値としては小さいものの、減少の幅はほぼ一致している。とすれば、受講した履修者の語彙は豊富になるのではなく、むしろ貧弱になっているようにみえる。この現象の意味、すなわち「語彙の貧弱さ」がいかなる要因に基づくのかを考える必要がある。

2020 年度・2021 年度ともに、授業内ではアカデミックワードの使用を推奨している。それぞれの年度のレポートの A 回と B 回を品詞別に比較すると、ほぼすべての分類において異なり語数が減少している。使用されている語の遷移をみると、日常的な語の使用が減少している。とすれば、この「語彙の貧弱さ」は、語の使用がアカデミックワードへと収斂していくためではないかと考えられる。さらに先稿では B 回以降において s の増加がみられることを確認している。よってレポート執筆に習熟した受講生は授業の進展にあわせて語彙が豊富になっていくことが予想される。

つまり、「国語表現」受講者においては、当初はレポートとして不適切な表現をしていたのが、適切な表現方法を獲得する過程において、見かけ上、語彙が貧弱になる。しかし、そこから新たな表現技術を獲得し、より適切な表現をすることが可能となり、再び語彙が豊富になる。このような成長過程を想定できるのではないか。

3.2　文章の体裁

つづいて、敬体の使用・段落の字下げ・段落数をみる。以下にそれらを数値化した一覧を示す。「敬体の使用」は、常体ではなく敬体を使用しているものの数である。「段落の字下げ」は、段落の冒頭について字下げをしていないものの数、「単段落」はレポート全体において段落数が 1 つのみのものの数を示す。

表 2　文章の体裁

	レポート総数	敬体の使用	段落の字下げ	単段落
2020A	87	20	41	11
2020B	84	3	13	1
2021A	84	18	49	12
2021B	81	2	20	3

表 2 にもとづいて、2021 年度の変化をみると、2020 年度と同様に肯定的な教育効果があったものと考えられる。数値の推移もほぼ同様の結果となっている。ただし、以下に示すように微細ながら傾向には差がみられる。

敬体の使用は、2020 年度は 20 件から 3 件に減少、2021 年度も 18 件から 2 件に減少している。ただし 2020B の 3 件は、いずれも 2020A では敬体でレポートを執筆していない履修者のもので、単純な減少ではない。一方、2021 年度の 2 件は、全体は常体で書かれているが、数文のみ、語尾が「～です」「～ます」とされているもので、実質的には 0 件といってよい。

段落の字下げは、2020 年度は 41 件から 13 件に減少し、2021 年度も 49 件から 20 件に減少している。2020 年度については単純な減少であるが、2021 年度は 2021A では字下げをしているにもかかわらず 2021B では字下げをしていない履修者のものが 2 件あった。

単段落、すなわち段落が 1 つのみのものは、2020 年度は 11 件から 1 件に減少、2021 年度も 12 件から 3 件に減少している。これも 2020 年度は単純な減少であるが、2021 年度はうち 1 件について、2021A では段落数が 8 であるのに対して 2021B は段落数が 1 となっているものがある。

以上をみると、文章の体裁については相当の教育効果を確認することができる。ただし、一部については否定的変化をみせるものがある。割合としては、全体の 5% 以内である。また、常体使用や適切な段落数の設定についての指導は高い効果があるのに比して、段落冒頭の字下げについての指導は効果が不十分である。受講生がなぜ段落冒頭の字下げを行わないのかについては、今後、アンケートの実施などによる原因究明を課題としたい。

3.3 作文能力

2021 年度に実施した「国語表現」では、先稿での分析結果に基づき、指導上の課題を設定して授業運営を行った。ここまでにみたとおり、特に指導をしやすい「文章の体裁」については肯定的な教育効果を確認することができた。

しかしながら、文章作成能力を向上させるための指導は容易ではない。稿者は比較的指導の難度が低い内容として、2020 年度「国語表現」の分析結果に基づき、2021 年度「国語表現」では適切な文長による文章作成とアカデミックワードの使用とについての指導を強化した。

まず、文長の確認から始めたい。文長の計測には MTmineR を使用し、一文の長さを 20 字ごとの階級に分け、201 字以上は同じ階級に分類されるように設定した。2021 年度授業では、200 字を超過する文でかつ意味をとりづらい

ものを例示し、内容の整理されていない長文を書かないように注意喚起した。授業内でも折に触れ、繰り返し適切な長さの文を書くように指導した。

以下の表3は20字ごとの文長を示したものである。

表3　文長

	s1-20	s21-40	s41-60	s61-80	s81-100	s101-120	s121-140	s141-160	s161-180	s181-200	s>=201
2020A	157	644	507	265	102	44	19	8	4	2	6
2020B	119	629	535	282	114	41	25	10	5	1	15
2021A	133	577	446	260	99	37	26	11	8	4	10
2021B	146	597	551	255	123	44	18	12	3	2	0

2020Aから2021Bまでの文長をサンプルとし、帰無仮説を「年度・時期に関係なく文長は変わらない」と設定して分散分析を行ったが、P値は0.998、F境界値は2.84であり、帰無仮説を却下するに至らなかった。つまり、統計的にはこれらのサンプルに有意な差がない。

ただし、2020年度と2021年度のレポートについて160字以上の文の数を比較すると、2020Aは12件、2020Bは21件と増加しているのに対し、2021Aは22件、2021Bは5件と減少した。さらに、2021Bでは201字以上の文は0件である。上述のとおり、2021年度授業では2020年度授業の反省に基づき、一文が長いとねじれ文になったり修飾が不分明になったりと、悪文になりがちであることについての指導を繰り返し行った。その効果があったものと考えられる。

次に副詞の使用状況をみる。ここで副詞に限定するのは、文脈の構成あるいは表象概念の構成を行う機能が小さいと考えられるためである。

頻度表の作成にはKH Coderを使用し、副詞および副詞B（ひらがなのみの副詞）を抽出し、比較する。2020Aから2021Bにかけて、使用頻度上位20位までの語をそれぞれ示す。

表４　副詞

2020A		2020B		2021A		2021B	
実際	33	実際	37	必ず	36	実際	32
必ずしも	19	必ず	22	実際	16	仮に	20
仮に	15	特に	18	改めて	14	最も	17
少し	15	既に	7	少し	13	特に	12
本当に	14	別に	7	特に	10	少し	10
最も	13	全く	6	当然	7	次に	6
当然	11	必ずしも	6	次に	6	当然	6
特に	11	仮に	4	最も	5	必ずしも	6
全く	8	極めて	4	決して	4	決して	4
比較的	8	最も	4	更に	4	同じく	4
共に	6	多々	4	同じく	4	ある程度	3
ある程度	5	当然	4	必ずしも	4	一概に	3
更に	5	未だ	4	仮に	3	一層	3
少なくとも	5	ある程度	3	何ら	3	益々	3
多少	5	一概に	3	元々	3	果たして	3
到底	5	果たして	3	今や	3	既に	3
一概に	4	少し	3	多少	3	更に	3
一見	4	同時に	3	同時に	3	多少	3
単に	4	もう一度	2	未だに	3	度々	3
果たして	3	何故	2	ある程度	2	同時に	3
決して	3	概ね	2	一概に	2	必ず	3
現に	3	現に	2	一気に	2	本当に	3
初めて	3	互いに	2	極めて	2		
徐々に	3	徐々に	2	現に	2		
大いに	3	度々	2	少なくとも	2		
同時に	3	同じく	2	常に	2		
		比較的	2	色々	2		
		本当に	2	生まれながら	2		
		未だに	2	長らく	2		
				未だ	2		

先稿でもみたとおりではあるが、副詞については大きな変動がみられない。下位の使用語をみると、2021B では「一層」「益々」「果たして」「本当に」など、

アカデミックワードとして適切ではない語の使用がみられる。2021 年度は 2020 年度実施授業の分析結果にもとづき、副詞の使用についての指導は配慮をしたが、受講生には徹底されなかった。使用頻度はいずれも 3 回とわずかではあるが、今後の課題したい。

表 5　副詞 B

2020A		2020B		2021A		2021B	
さらに	51	さらに	35	まず	39	さらに	42
まず	48	どう	29	なぜ	23	まず	23
もし	25	まず	27	さらに	22	より	17
どう	24	ほとんど	16	どう	19	なぜ	15
そう	21	そう	10	ほとんど	16	すでに	13
あまり	17	なぜ	10	より	13	どう	12
なぜ	13	あくまで	8	もし	12	そう	9
やはり	12	より	7	あまり	9	むしろ	9
これから	11	これから	6	そう	8	いずれ	8
より	11	もちろん	6	もっと	8	かつて	7
また	10	むしろ	5	そのまま	7	もし	6
かなり	9	もし	5	もちろん	7	かえって	5
ますます	9	わずか	5	よく	7	これから	5
むしろ	8	あまり	4	こう	6	あまり	4
もちろん	8	あまりに	4	これから	6	ほぼ	4
すでに	7	いまだ	4	ほぼ	6	やはり	4
ほとんど	7	しっかり	4	まだまだ	6	いかに	3
もう	7	そのまま	4	あくまでも	5	およそ	3
かつて	6	たしかに	4	あらかじめ	5	たしかに	3
もっと	6	もう	4	たしかに	5	たとえ	3
あまりに	5	あくまでも	3	やはり	5	ほとんど	3
こう	5	あらかじめ	3	わずか	5	ますます	3
このように	5	こう	3			まだ	3
そもそも	5	さほど	3			もう	3
たしかに	5	そもそも	3			もしか	3
たとえ	5	とても	3			もちろん	3
とても	5	ほぼ	3				
		まだ	3				
		やはり	3				

副詞 B については、2021B に変化が観測される。2021A では「もっと」(8 例)「そのまま」(7 例)「もちろん」(7 例)「よく」(7 例)「こう」(6 例) などの使用があるのに対し、2021B ではみられない。一方、2021B では「すでに」(13 例)「いずれ」(8 例)「かつて」(7 例)「かえって」(5 例) など、2021A ではみられない語の使用が確認できる。2021A ではアカデミックワードとして適切でない語が使用されていたのが 2021B では使用されなくなり、2021B では適切である語が使用されている。

以上をみるに、2021 年度「国語表現」では、副詞の使用については指導上の課題を残すものの、副詞 B については教育効果を認めることが可能である。

4. 論理構成

4.1 対応分析

ここまでは先稿での研究成果をもとに、2020 年度「国語表現」と 2021 年度「国語表現」とを対照し、分析を行った。本節および次節では、2 種のレポートテーマについて論理構成の面から分析を行う。先に示したとおり、2021A と 2020B では「選択的夫婦別姓」導入の是非についてのレポートを課し、2020A と 2021B では「独身税」導入の是非についてのレポートを課した。前者は「是」の立場でのものが大半であり、後者は「否」とするものが大勢を占める。つまり、いずれのテーマも結論の方向性は同一である。しかし、論理構成はレポート執筆時期に関係なく同一であろうか。仮に、レポート執筆に習熟したときに、体裁や表現のみならず論理構成についても能力が向上しているのであれば、2021A と 2020B あるいは 2020A と 2021B では異なった様相をみせるはずである。提出されたレポートからは、一定期間を経たのちに執筆されたものの方が論理構成も巧みになっている印象をうける。しかしながら、これはどこまでも「印象」である。

この「印象」を検証する目的から、本稿ではレポート執筆時期による論理構成の変化の観測を試みる。そのため、KH Coder の対応分析機能および階層的クラスター分析機能を使用して、それぞれの課題の論理構成を探索する。

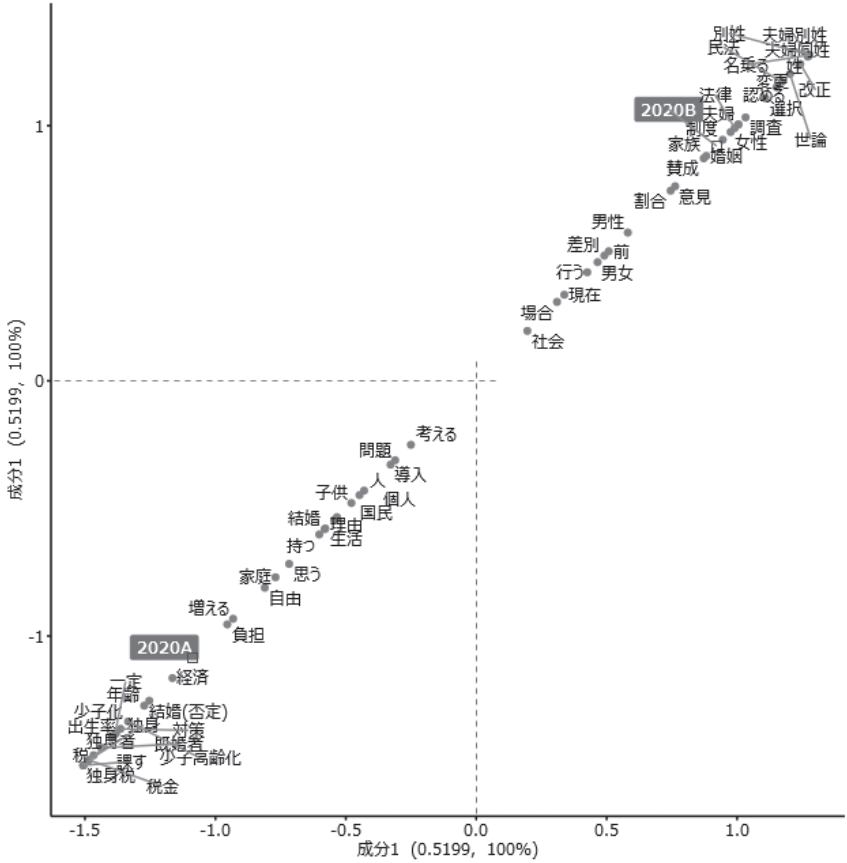
まず、すべてのレポートについての対応分析を行う。変数として 2020A・2020B・2021A・2021B を設定する。

対応分析とは「頻度表における行・列の関係を組み替え，頻度表に含まれる情報を少数の成分（次元）にまとめることで，行･列を整理する解析法である」(石川慎一郎･前田忠彦･山崎誠:2010)。KH Coder による対応分析については、

樋口耕一（2019）に詳しい。

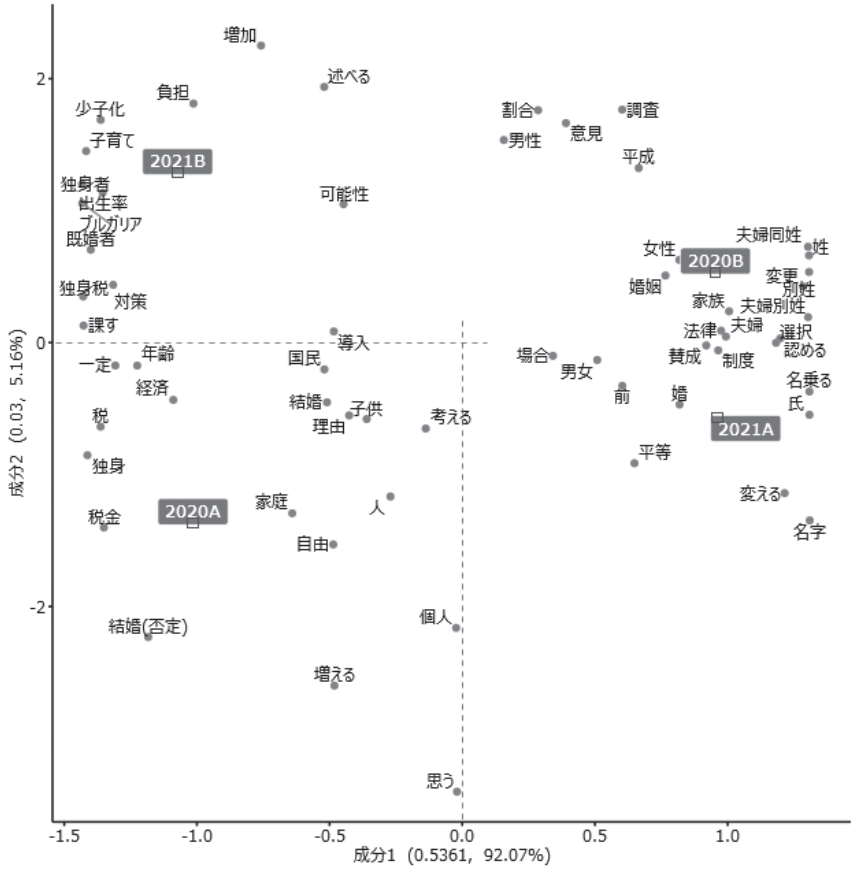
以下は、変数を 2020A と 2020B として対応分析を行った結果である。変数が 2 種のみであり、かつテーマが異なるため、線状に分布している。

図 1　2020A と 2020B の対応分析



仮に 2021A と 2020B、また 2020A と 2021B とに内容面での変化がなければ、この図 1 のように二極化した分布をみせるはずである。しかし、以下の図 2 にみるように、2021A・2020B・2020A・2021B は四象限に分布する。

図２　2021A・2020B・2020A・2021B の対応分析



上図は本稿で分析対象とするすべてのレポートについての対応分析を行った結果である。変数の傾向をみると、右上方向に 2020B、左上方向に 2021B、左下方向に 2020A、右下方向に 2021A が位置している。成分１（左右）を軸とすれば、左方向が「独身税」についてのレポート、右方向が「選択的夫婦別姓」についてのレポートとみることができる。また、成分２（上下）を軸とすれば、下方向は受講直後のレポート、上方向は受講から一定期間が過ぎてからのレポートとみることが可能である。

成分２に注視すると、2020B と 2021A は比較的近い位置にあるのに対して、2020A と 2021B とには距離がある。これは、「選択的夫婦別姓」についてのレポート（2020B および 2021A）は受講時期に関わりなく内容が類似していることを

示唆する一方で、「独身税」についてのレポート（2020A および 2021B）は受講時期によって内容にちがいがみられることを示している。

なお、成分 1 の寄与率が 92.07% となっているため、対象となるサンプルの特徴の多くは左右の軸、つまりレポートテーマの差として説明される。一方、成分 2 の寄与率は 5.16% と、成分 1 に比して低い。しかしながら、図 1 では成分 1 のみ、つまりレポートテーマの差のみで寄与率が 100% となっていたことに鑑みれば、成分 2 の寄与率は無視するべきものではないと考えられる。

とはいえ、図 2 から看取することができるのは、あくまで「4 種の変数に傾向がみられる」ことのみである。その傾向がいかなる意味をもつのかは、より適切な分析を経て明らかになる。そのため、それぞれのレポートがいかなる論理構成を成すのか、次節では階層的クラスター分析を行い、レポートの内容について考える。

4.2　階層的クラスター分析

階層的クラスター分析とは「データが持つ情報を手掛かりにして、距離の近いデータ同士をまとめてクラスター（群, 集落）を形成する統計手法である」（石川他：2010）。KH Coder における抽出語の階層的クラスター分析は、「出現パターンの似通った語の組み合わせにはどんなものがあったのか」（樋口：2020）をみるものである。分析対象となるテキストを形態素解析し、出現パターンに類似性のある語をデンドログラムで表示する。そして、任意の併合水準でグルーピングすることで、語の群の特徴をみる。

なお、階層的クラスター分析を行うにあたっては石川他（2010）および樋口（2020）を参考に、距離測定法としてウォード法を、距離としてユークリッド距離を使用する。また、それぞれのレポートについて、対象となる語が 35 件となるように調整した。各サンプルにおける語の最小出現数は、2020A が 63 件、2020B が 70 件、2021A が 71 件、2021B が 74 件である。

図３　2021A 階層的クラスター分析

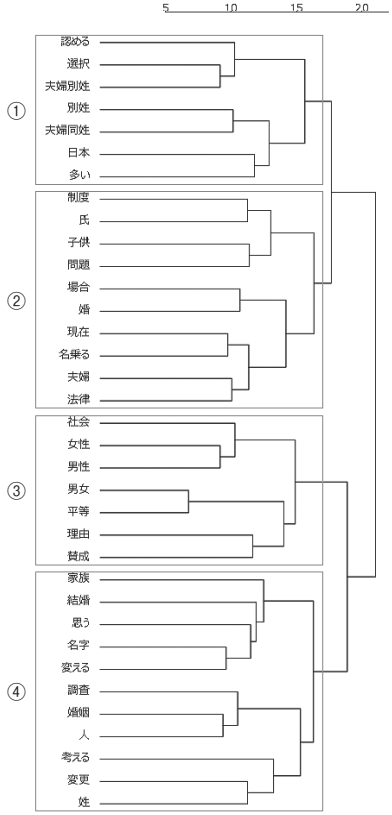
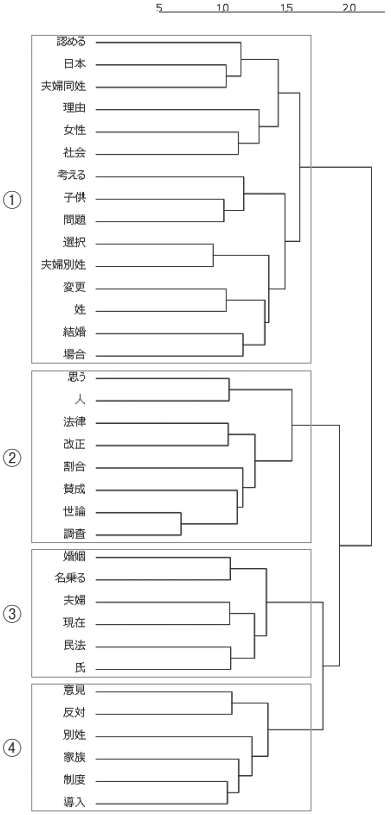


図４　2020B 階層的クラスター分析



2021A および 2020B の階層的クラスター分析の結果を確認し、比較する。2021A・2020B ともに併合水準を 4 に設定した。

2021A の内容をみると、「日本の状況確認と選択的夫婦別姓導入」「夫婦別姓となった場合の子供の帰属問題」「女性の立場に鑑みた当該制度への賛成意見」「婚姻に伴う姓変更についての意識調査」とまとめられる。

①では「認める」「選択」「夫婦別姓」と「別姓」「夫婦同姓」「日本」「多い」がグループになっている。選択的夫婦別姓の導入を認めることと日本では夫婦同姓により多くの問題が生じている状況とが結びついている。②は「制度」「氏」「子供」「問題」と「場合」「婚」「現在」「名乗る」「夫婦」「法律」とがグループを構成している。夫婦の氏（姓）が異なる場合に子供の帰属が問題となること、現在は法律によって名乗る姓が固定されていることとが論じられている。

つまり、現在は法制度によって夫婦の姓が固定されているが、別姓となった場合に子供の姓はどちらとなるかが問題になると指摘されている。③では「社会」「女性」「男性」「男女」「平等」「理由」「賛成」より、女性の社会的状況や男女平等の観点から当該制度導入への賛意が主張されている。④は「家族」「結婚」「思う」「名字」「変える」と「調査」「婚姻」「人」「考える」「変更」「姓」より、婚姻による姓の変更についての調査結果への言及と解釈できる。

2020B の内容をまとめると、「〈夫婦同姓が女性の社会進出を阻害していること〉と〈夫婦別姓が子供にとって問題となること〉との対比」「夫婦別姓を支持する世論調査の結果」「民法による規定の確認」「当該制度導入に伴う家族制度への悪影響を懸念する反対意見」である。

①は「夫婦同姓」と「夫婦別姓」を含む 2 つのツリーで構成されている。前者は「女性」「社会」「理由」などから、夫婦同姓が女性の社会進出において阻害条件となっていることを述べる。後者は「結婚」「子供」「問題」などから、結婚後に夫婦別姓となった場合に子供の姓が問題となることを述べる。②は世論調査の結果、法律改正を望む人の割合が高いことを論拠として提示している。③は「現在」「民法」「氏」等から、現行の民法における婚姻の際に夫婦がどちらか片方の氏を名乗ることについての規定を確認している。④は「制度」「導入」「家族」「別姓」「反対」「意見」から、当該制度が導入されると家族制度そのものに悪影響を与えかねないとする反対意見を紹介している。

2021A と 2020B とを比較すると、類似点が認められる。いずれも、当該制度導入の論拠として意識調査あるいは世論調査の結果を提示しており、女性の社会進出の観点から賛意を示す。また、双方ともに想定される反対意見として、子供の帰属が挙げられている。

一方、2021A と 2020B との相違としては、前者に対して後者がやや複雑な論理構成をとっていることを指摘しうる。2021A は上記の類似点とあわせて日本の状況確認をしているのみであるのに対し、2020B は当該制度導入に対する反対意見として家族制度への悪影響を指摘している。また、夫婦同姓と夫婦別姓とを対比的に論じている。

以上をみるに、2021A と 2020B とは内容に相当の類似性を認められるが、2020B の方がやや高度な論理構成を行っているといえる。

続いて、2020A と 2021B とについて、階層的クラスター分析にもとづいた検討を行う。いずれも併合水準は 6 に設定した。

図５　2020A 階層的クラスター分析

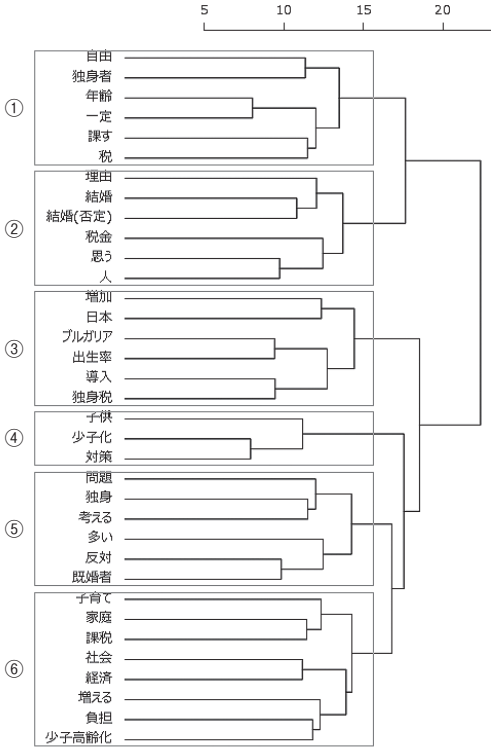
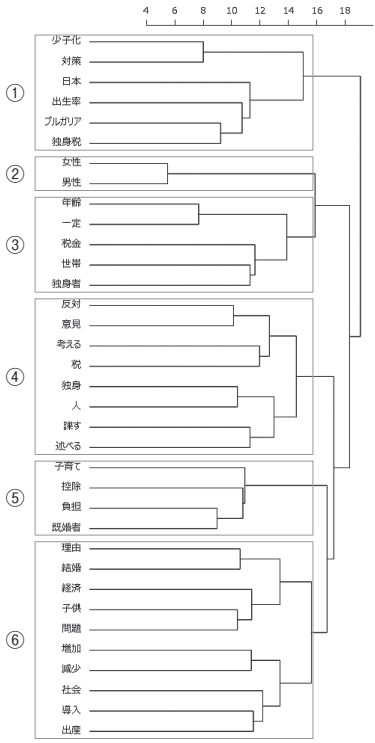


図６　2021B 階層的クラスター分析



2020A は「独身税の導入に伴って独身者の自由が阻害される問題」「婚姻を課税によって制限することへの不満あるいは忌避の予測」「ブルガリアの事例と日本」「少子化対策」「独身税が既婚者の増加に寄与しないこと」「子育て支援政策の必要性と少子高齢化社会の問題」とまとめることができる。

①は「自由」「独身者」「年齢」「一定」「課す」「税」から、一定年齢の独身者に課す税（独身税）が独身者の自由を阻害するものであると述べている。②は「理由」「結婚」「結婚（否定）」「税金」「思う」「人」で構成されている。「思う」がレポートの書き手の意思の表示であるのに対し、「人」は「独身の人」のように書き手以外が想定されている。とすれば、このグループは結婚するあるいは結婚しない理由の推測と、税金（独身税）を課すことに対して不満を覚える人がいると書き手が思っていることを示しているとみられる。③は「増加」「日本」「ブルガリア」「出生率」「導入」「独身税」から、かつて出生率の向上のためにブルガリアで導入された独身税の事例提示と日本の状況が対比されて

いるものと考えられる。④は「子供」「少子化」「対策」から、独身税に少子化対策の面があることを示す。⑤は「問題」「独身」「考える」「多い」「反対」「既婚者」から、独身（者）と既婚者を対比し、独身税の導入が既婚者の増加には寄与しないと問題提起しているものとみられる。⑥は「子育て」「家庭」「課税」「社会」「経済」「増える」「負担」「少子高齢化」より、家庭における子育ての補助は独身税とは別の手段で行うべきこと、少子高齢化による負担増が社会と経済に悪影響を及ぼすことについて述べている。

2021B は「ブルガリアの事例と日本」「男女の比較データ」「独身税の是非」「独身税への反対意見」「子育て世帯への控除の必要性」「経済と子供が結婚を左右する事実」「独身税が導入された場合の社会への影響」とまとめられる。

①は「少子化」「対策」「日本」「出生率」「ブルガリア」「独身税」より、2020A と同様にブルガリアの事例と日本とを比較している。②では「女性」と「男性」が対比されている。対比されている内容としては未婚率・離婚率・結婚理由・年収等が確認される。③は「年齢」「一定」「税金」「世帯」「独身者」から一定年齢の独身者に課す税金（独身税）を単身（独身）世帯に課すことの是非を論じている。④は「反対」「意見」「考える」「税」「独身」「人」「課す」「述べる」より、独身税に対する反対意見の表明と読み取れる。⑤は「子育て」「控除」「負担」「既婚者」から、独身税を課さない場合に子育てをしている既婚者への負担を軽減するために、何らかの控除を導入する必要性を述べている。⑥は「理由」「結婚」「経済」「子供」「問題」と「増加」「減少」「社会」「導入」「出産」より、結婚の理由が経済と子供の問題に左右されること、また独身税の導入が結婚数の増加にはつながっても出生率減少の歯止めにはならず、社会に対して肯定的影響を与える可能性が低いことについて論じている。

2020A と 2021B の内容を比較すると、いくつかの類似点がみられる。まず、いずれもブルガリアの事例を紹介している。これは独身税をテーマにしたものに限らない、きわめて一般的な事実であるが、受講生に事前知識のなさそうな課題内容のレポートを課すと、彼らはウェブ検索を行い、上位に提示された内容をそのまま用いる傾向がある。事実、「独身税」でウェブ検索を行うと、上位にくるのはブルガリアの事例紹介である。なお、「ブルガリアで独身税が導入されたことがあるようです」という、一次資料が不明瞭な検索結果が大半ではある。次の類似点として、独身税とは異なる子育て支援についての言及が指摘できる。ただし、2020A が漠然とした内容であるのに対し、2021B では「控除」としている。

それらの類似点に対し、相違する点が多くみられる。2020A では「独身税

による独身者の自由の阻害」「婚姻を課税によって制限することへの不満あるいは忌避」「独身税が既婚者の増加に寄与しないこと」が述べられているのに対し、2021B では「男女の比較データ」「独身税の是非」「独身税への反対意見」「経済と子供が結婚を左右する事実」「独身税が導入された場合の社会への影響」が論じられる。2020A の論理構成は独身税への反対意見が軸になっているのに対し、2021B の論理構成は問題の設定と主張の提示と同時に、独身税を取り巻く状況分析が多く行われている。

以上に鑑みるに、2020A と 2021B とには類似よりも相違の傾向が強い。また、2021B は問題提起・主張・分析・論証がより丁寧であるとみることが可能である。

前節では 2021A と 2020B との差は小さく、一方で 2020A と 2021B とでは差が大きいことをみた。本節の結果で行った分析においても、類似の傾向を確認した。

おわりに

最後に、ここまで述べきたったことを振り返り、まとめとしたい。

本稿の目的は、2020 年度および 2021 年度に稿者が担当した「国語表現」において提出されたレポートを対象に、学習効果を観測することであった。そのため、先稿（植田:2020）において得られた成果を基礎として、「語彙の豊富さ」「文章の体裁」「作文能力」についての観測を行い、さらに「論理構成」についての検討を行った。

「語彙の豊富さ」「文章の体裁」「作文能力」については、2021 年度においても 2020 年度と同様、あるいはそれ以上の結果がみられた。すなわち、いずれも授業受講当初に比して一定期間を経過したあとのレポートに肯定的結果を観測することができた。

また、本稿ではテキストマイニングによるレポート課題の論理構成分析を試みた。手法としては、対応分析による全体的な傾向確認と階層的クラスター分析による個別の論理構成分析である。対応分析によって、レポート内容および執筆時期に傾向があることを確認した。さらに、階層的クラスター分析によって対応分析で得られた結論をより詳細に分析した。この考察により、レポート内容の論理構成における分析手法として、テキストマイニングが有効であることが明らかとなった。

先稿においては、テキストマイニング技術が教育効果の測定およびレポート

採点の際の補助として運用可能であることを確認した。本稿においても、先稿の結論を補強できた。また、対応分析および階層的クラスター分析がレポートの内容あるいは論理構成の分析に耐えることも確認された。

補注

本稿で用いた技術については先稿に詳しいが、2021 年度授業では使用したレポート用フォーマットを改訂しており、本稿においても説明を要するところがある。そのため、技術面での概略を以下に示す。

通常、テキストマイニングを行う際、分析対象とするファイル形式は、テキストファイルもしくは csv ファイルである。しかしながら、学生にレポートを課すとき、それらのファイル形式を指定するケースは稀であろう。そのため、稿者はウェブのアンケートフォームを模したエクセル用のファイルを設計し、それをレポート用のフォーマットとして受講生に提示した。

このフォーマットでは、先稿（2020 年度）の時点では「学生番号」「氏名」「レポート本文」を記入させ、

①学生番号をファイル名とし、レポート本文をテキストとしたテキストファイル

②学生番号・氏名・レポート本文をそれぞれの列におさめる csv ファイル

の 2 種を作成するようにシステム設計した。2021 年度授業で用いた改訂版フォーマットでは「学生番号」「氏名」「レポート本文」に加え、「レポートタイトル」「テーマ記号」「参考文献等」を記入する欄を設定した。ただし、出力するデータは先稿で用いたものと同様である。

①のテキストファイルは MTmineR での分析に、②の csv ファイルは KH Coder での分析に使用した。

段落数・文長については、MTmineR を使用して計測した。MTmineR を使用するためには各種のアプリケーションソフトにパスを通す必要があるものの、段落数のように単純に計算するだけのものであれば、本体と JRE（Java Runtime Environment）の導入だけで済む。

敬体の使用・段落の字下げについては、csv ファイルで確認した。敬体はレポート本文の列で「です」「ます」を含むものを抽出した。段下げは、レポートを csv ファイルに変換する際に、レポート本文の 1 文字目に空白スペースが含まれるか否かを機械的に判別し、含まれるものを「1」、含まれないものを「0」として判定するようにシステムを設計した。

延べ語数と異なり語数の計測、頻度表の作成、また対応分析と階層的クラス

ター分析にはKH Coderを使用した。なお、KH Coderでは通常、助詞・助動詞等は分析対象としないため、延べ語数・異なり語数にも含まれない。

引用文献

石川慎一郎・前田忠彦・山崎誠（2010）:『言語研究のための統計入門』くろしお出版

植田麦（2020）:「テキストマイニング技術を応用したレポート課題の教育効果測定」『實踐國文學』99

鄭弯弯・金明哲（2018）:「変動係数を用いた語彙の豊富さ指標の比較評価」『同志社大学ハリス理化学研究報告』58（4）

樋口耕一（2019）:「計量テキスト分析における対応分析の活用　―同時布置の仕組みと読み取り方を中心に―」『コンピュータ＆エデュケーション』47

樋口耕一（2020）:『社会調査のための計量テキスト分析　―内容分析の継承と発展を目指して（第2版）』ナカニシヤ出版

（うえだ　ばく・明治大学准教授）