2024 年度　統計データ分析コンペティション

審査員奨励賞　[大学生・一般の部]

中学生の言語による表現を巡る規定要因分析
―潜在意味解析と Elastic Net 回帰を用いた分析―

陣内 未来（九州大学大学院人間環境学府）、
立山 皓基（九州大学教育学部）

中学生の言語による表現を巡る規定要因分析
—潜在意味解析と Elastic Net 回帰を用いた分析—

陣内未来*1・立山皓基*2

*1：九州大学大学院人間環境学府
*2：九州大学教育学部

１．序論：言語による表現に対する規定要因

教育において、子どもの「言語力」(注 1)の育成は国内外の多方面から高い関心を集めている。例えば国内においては、現行の学習指導要領では言語能力を「学習の基盤」と位置付けており、高い重要度が示されている(文部科学省, 2017)[7]。また国際的にも、経済協力開発機構（OECD）が実施する国際学力調査である PISA 調査では 2000 年から最新の 2022 年調査まで「読解リテラシー」に関する学力調査と、それに関する質問紙調査が継続的に実施されており(文部科学省・国立教育政策研究所, 2023)[8]、やはり高い注目を集めている。もちろん、学習指導要領における「言語能力」や PISA 調査における「読解リテラシー」に限らず、各方面で想定されている「言語力」は互いに異なる概念であるが、国の内外を問わず、子どもの「言語力」の育成に対する高い関心が寄せられていることは間違いないだろう。

ところで、このような高い関心を集める「言語力」は何によってもたらされるのだろうか。人は生まれてから様々な場面で言語に触れていくことで、言語を獲得していく。Hoff(2016)[4]によれば、そのような言語発達(Language Development)には、遺伝要因と環境要因の 2 つが働いているとされる。その上で、言語発達に対する遺伝要因の占める分散は 1~82%と非常に広いため、言語発達の理解には社会的要因が重要になる。社会的要因に着目すると、そもそも言語発達には「相互コミュニケーションの相手」と「モデルとなる言語の存在」（日本であれば日本語となる）が必要不可欠であり、個人が育つ文化や社会経済的要因などは前者の質と量を規定し、民族性などの要因は後者を規定する。言語発達の分散に対してこれらの社会的要因の説明率を具体的に言及することは難しいものの、これらの変数によって言語発達に差が出ることが指摘されている（Hoff 2006）[4]。

Hoff による網羅的レビューは様々な国・社会を対象とした研究群を参照しているが、日本国内でも子どもの「言語力」に対する社会的な規定要因を明らかにする試みは様々行われてきた。しかし、そのほとんどは国語などのテスト得点を従属変数にしており(例えば、鳶島, 2016[11]; 志水ほか, 2014[10]; 山田, 2014[12]など)、実際に個人が産出した言語を対象とはしていない。そのため、文法や漢字、読解力などテストで測定しやすい側面については理解が進みつつあるが、言語による表現のようなテストで測定しにくい側面についてはほとんど触れられてこなかった。一方で、教育社会学者の渡邉雅子によれば、日本の作文教育においては、「日常生活で強く印象に残った出来事の一場面を描写し、そこから何を感じ考えたか、書き手が体験から学んだことを書く。出来事の生き生きとした描写と気持ちの表現、結論として書き手の精神的な成長がうかがえる今後の行為の指針や心構えが書けている」(渡邉, 2023, p.227)[15]ことが求められ、日本の特徴であるという。この渡邉の指摘を踏まえるならば、当然、児童生徒がどのような日常生活を送っているか、すなわち児童生徒を取り巻く社会経済環境や自然環境、人的環境が彼らの表現に大きな影響を与えると考えられる。この指摘は渡邉によって度々なされてきた(渡邉, 2001[13]; 渡邉, 2007[14]など)が、生の作文データと書き手を取り巻く様々な環境変数を結び付けた分析が求められる分析上の困難さも相まってか、これまで統計的な実証研究に裏付けられてこなかった。

以上を踏まえて、本研究では子どもの「言語力」の内、その表現に着目し、彼らの繰り出す表現の規定要因分析を行う。具体的には、中学生が言語によって繰り出す表現が何によってもたらされるのかを、地域の社会経済環境や自然環境、人的環境に着目して明らかにする。

２．分析の枠組み：潜在意味解析と Elastic Net 回帰を用いた表現の規定要因分析の試み

本研究では、後述する作文データにおける表現を従属変数にし、個人が顕在的・潜在的に接触しうる地域の社会経済環境、自然環境、人的環境を独立変数とする。そのため、まず各作文の表現の抽出法として、本研究では潜在意味解析(LSA: Latent Semantic Analysis)を用いる。潜在意味解析はテキストデータの解析に使われる手法の一つであり、文書と単語の間の潜在的な意味関係をモデル化するために使われる（Deerwester 1990）[2]。潜在意味解析では、高次元の単語-文書行列から低次元の「意味」空間(“semantic” space)を構築し、一般的に「トピック」と呼ばれる潜在的な共起性を見出す。具体的には、式（1）のように単語-文書行列 A（各要素は TF-IDF による重みづけを行っている）に対して特異値分解を適用することで表現される。

$$A = USV^T \quad (1)$$

ここで、Uは単語-トピック行列、Sは特異値を対角成分とする行列、V^Tは文書-トピック行列の転置行列である。この時、上位 k 個の特異値(および対応する特異ベクトル)を選択し、次元削減を行う。即ち、ランク k の近似行列を A_k、上位 k 個の左特異ベクトルによる行列を U_k、上位 k 個の特異値を含む対角行列を S_k、上位 k 個の右特異ベクトルによる行列の転置行列を V_k^T とした時、元の単語-文書行列 A の近似は式(２)のように表現できる。

$$A_k = U_k S_k V_k^T \quad (2)$$

ここで、U_kはある次元（トピック）に対する単語の、V_k^Tはある次元に対する文書の関連性の強さを示す重みベクトルと言える（Deenwester 1990, p.395）[2]。そこで、U_kからある次元に対する単語の対応関係を検討することで、このトピックの解釈を行うことができる。

その上で、V_k^Tの各行は各列に位置する文書があるトピックに対して有する重みを表現しているため、これを従属変数とし、個人が接触しうる地域社会の諸変数を独立変数にして、Elastic Net 回帰 (Zou & Hastie, 2005)[16]による解析を行う。これにより、あるトピックの表現に対する規定要因を検討することができる。

なお、本研究では独立変数に社会経済環境変数を用いているが、一般的に社会経済環境変数は変数間の相関が高くなり、通常の重回帰分析等では一度に扱うことができない。そこで、この問題に対して Elastic Net 回帰を用いることで、変数間の相関が高い変数を同時に扱いつつ、変数選択を行うことが提案されている(荒川・野寄, 2023)[1]。具体的に、$\boldsymbol{y}$ を従属変数、n をサンプルサイズ、p を変数の数、$n \times p$ の行列 $\boldsymbol{X}$ を独立変数、$\boldsymbol{\beta}$ を係数、$\boldsymbol{\alpha}$ と $\boldsymbol{\lambda}$ をハイパーパラメータとしたとき、Elastic Net 回帰は式(３)のパラメータ $\boldsymbol{\beta}$ に対する最適化問題を解くことで求められる(川野ほか, 2018, p.30[5])。Elastic Net 回帰式では通常の線形回帰式に対して L1 正則化項と L2 正則化項が加えられていることで、多重共線性の恐れがあるような互いの相関が高い独立変数を同時に回帰式に含むことが可能となっている(Zou & Hastie, 2005[16]; 川野ほか, 2018, p.30[5])。

$$L = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \boldsymbol{\lambda}\sum_{j=1}^{p}\left\{\alpha|\beta_j| + \frac{(1-\alpha)\beta_j^2}{2}\right\} \quad (3)$$

なお、ハイパーパラメータである α と λ については、$0.01 \leq \alpha \leq 0.99$ の範囲内において 0.01 刻みでグリッドサーチを行うと共に、5-fold クロスバリデーションによって決定する。

以上の手順を踏まえることで、個人が顕在的・潜在的に接触し得る地域の社会経済環境、自然環境、人的環境が個人の表現をいかに規定するかを検討することができる。

３．データ

本研究では、国土交通省が中学生を対象に毎年行っている「全日本中学生水の作文コンクール」[6]の受賞作品を基にデータを構築した。本データを用いる利点として第一に、中学生を対象としたコンクールの受賞作品である点が挙げられる。言語データの分散については、表現による分散のみならず、文法や語彙に基づく言語運用能力による分散によっても説明されると考えられる。そこで、ある程度の言語運用能力を身に着けている中学生、かつ本コンクール受賞作品を検討対象に据えることで、言語運用能力を統制しつつ、表現に着目した

分析が可能になると判断した。第二に、本コンクールの課題は毎年「水について考える」で統一されており、表現に高い自由度が認められている。これにより、ある程度共通の話題について書かれつつ、適度な分散が確保されると考えた。また、第三に、本コンクールの受賞作品には都道府県と学校名が記されているため、都道府県・市区町村別のデータと結合できる利点がある(注2)。第四に、毎年約 40 本の受賞作品が 2004 年から 2023 年まで国土交通省の WEB サイト[6]上で公開されており、分析に耐えうる量が確保できると判断した。本研究の検討対象は海外の日本語学校出身者の作品を除いた国内の中学生 715 名の作文を対象とした。

データセットは受賞作品データと文部科学省の学校コード、SSDSE（教育用標準データセット）、e-Stat 等より独自に確保した国の基幹統計などを結合し、構築した。使用する独立変数を表 1 に示す。まず、統制変数として学校の設置主体（国立私立:1、公立:0）を入れた。次に、地域の人的環境変数として核家族世帯数、母子世帯数、父子世帯数、60 歳以上人口、農林漁業者、農林漁業雇用者、労務作業者、会社団体役員、教員・宗教家、文筆家・芸術家・芸能家を投入した。なお、地域の人的環境変数の内の各職業については児童の「水の作文」に効果を及ぼすと考えられる職業に限定している。農林漁業者と農林漁業雇用者は児童に水に関する体験を与える可能性を考慮している。労務作業者と会社団体役員、教員・宗教家、文筆家・芸術家・芸能家は、これらの職業従事者が児童に対して与え得る教育水準の影響や文化的影響を考慮している。地域の社会経済環境変数として、中学校 1 校当たりの教員数、1 校当たりの生徒数、教育費、中学校費、社会教育費、図書館数を投入した。最後に、地域の自然環境変数として平均気温、日最高気温の平均、日照時間の合計、降水量の合計、降雪量の合計、主要湖沼面積、一級河川延長、二級河川延長、準用河川延長を投入した。

これらの変数の概要を表 1 に、要約統計量を表 2 に示す。

表 1．変数の概要

No.	変数	単位	集計単位	データの拠出
1	国立私立ダミー			文部科学省 HP 「学校コード」 https://www.mext.go.jp/b_menu/toukei/mext_01087.html
2	核家族世帯数	世帯	市町村	住民基本台帳人口要覧（e-Stat：A810102）
3	母子世帯数	世帯	市町村	国勢調査（e-Stat：A8401）
4	父子世帯数	世帯	市町村	国勢調査（e-Stat：A8501）
5	60 歳以上人口	人	市町村	国勢調査、人口推計（e-Stat：A1416）
6	農林漁業者	人	市町村	就業構造基本調査（e-Stat：F1301）
7	農林漁業雇用者	人	市町村	就業構造基本調査（e-Stat：F1302）
8	労務作業者	人	市町村	就業構造基本調査（e-Stat：F1303）
9	会社団体役員	人	市町村	就業構造基本調査（e-Stat：F1309）
10	教員・宗教家	人	市町村	就業構造基本調査（e-Stat：F1310）
11	文筆家・芸術家・芸能家	人	市町村	就業構造基本調査（e-Stat：F1315）
12	1 校当たり教員数	人/校	市町村	学校基本調査（e-Stat：E3401, E3101）
13	1 校当たり生徒数	人/校	市町村	学校基本調査（e-Stat：E3501, E3101）
14	教育費	千円	市町村	地方財政状況調査（e-Stat：D320310）
15	中学校費	千円	市町村	地方財政状況調査（e-Stat：D3203103）
16	社会教育費	千円	市町村	地方財政状況調査（e-Stat：D3203107）
17	図書館数	館	市町村	社会教育調査（e-Stat：G1401）
18	一級河川延長	Km	都道府県	国土交通省 HP「一級河川等の河川延長調*都道府県別（令和 5 年 4 月 30 日現在）」https://www.mlit.go.jp /statistics/details/content/001594706.pdf
19	二級河川延長	Km	都道府県	
20	準用河川延長	Km	都道府県	
21	主要湖沼面積	ha	都道府県	SSDSE-F-2023
22	平均気温	℃	都道府県	SSDSE-F-2023
23	日最高気温平均	℃	都道府県	SSDSE-F-2023
24	日照時間の合計	時間	都道府県	SSDSE-F-2023
25	降水量の合計	Mm	都道府県	SSDSE-F-2023
26	降雪量の合計	Cm	都道府県	SSDSE-F-2023

注 1．集計単位について、特別区については「区」を単位として集計している。
注 2．「1 校当たり教員数」「1 校当たり生徒数」は学校基本調査における教員数と生徒数をそれぞれ中学校数(e-Stat：E3101)で除して求めている。

表２．要約統計量

	n	平均値	標準偏差	最小値	最大値
1．国立私立ダミー	715	0.3	0.5	0	1
2．核家族世帯数	715	80631.1	122811.8	92	971451
3．母子世帯数	715	2034.8	2850.6	2	24184
4．父子世帯数	715	214.8	296.4	0	2546
5．60 歳以上人口	715	101650.4	143434.9	157	1120861
6．農林漁業者	715	2967.5	2794.2	0	17200
7．農林漁業雇用者	715	493.5	474.8	0	2520
8．労務作業者	715	10574.6	14702.9	30	147432
9．会社団体役員	715	3114.3	4908.7	0	41450
10. 教員・宗教家	715	7111.2	9597.3	20	79110
11. 文筆家・芸術家・芸能家	715	2708.8	5043.5	0	42590
12. 1 校当たり教員数	715	24.1	6.6	4	42
13. 1 校当たり生徒数	715	327.5	136.1	4	673
14. 教育費	715	19273436.6	36873710.8	78468	301873300
15. 中学校費	715	3203029.7	7388271.1	7165	55865180
16. 社会教育費	715	2641478.8	3686332.3	1616	26931080
17. 図書館数	715	5.4	5.7	0	26
18. 一級河川延長	715	1765.6	1498	0	10189
19. 二級河川延長	715	765.5	737.3	0	4287
20. 準用河川延長	715	456.1	451	33	1966
21. 主要湖沼面積	715	236.2	1281.9	0	14307
22. 平均気温	715	15.8	2.4	9	23
23. 日最高気温平均	715	20.3	2.3	13	26
24. 日照時間の合計	715	1941.4	168	1527	2226
25. 降水量の合計	715	1645.4	407.3	965	2666
26. 降雪量の合計	715	56	118.8	0	567

表３．データ間の年度対応

<table>
<tr><th rowspan="2">作文</th><th colspan="26">独立変数（各変数の番号は表２に対応）</th></tr>
<tr><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th><th>16</th><th>17</th><th>18</th><th>19</th><th>20</th><th>21</th><th>22</th><th>23</th><th>24</th><th>25</th><th>26</th></tr>
<tr><td>2023</td><td colspan="10" rowspan="4">2020</td><td rowspan="2">2022</td><td rowspan="4">2019</td><td rowspan="3">2018</td><td rowspan="20">2023</td><td rowspan="3">2021</td><td colspan="5" rowspan="20">1991 年～2020 年の
30 年間の平均値
（2020 年平年値）</td></tr>
<tr><td>2022</td></tr>
<tr><td>2021</td><td>2021</td></tr>
<tr><td>2020</td><td>2020</td></tr>
<tr><td>2019</td><td colspan="10" rowspan="4">2015</td><td>2019</td><td>2019</td></tr>
<tr><td>2018</td><td>2018</td><td>2018</td><td>2018</td></tr>
<tr><td>2017</td><td>2017</td><td>2017</td><td rowspan="3">2015</td><td>2017</td></tr>
<tr><td>2016</td><td>2016</td><td>2016</td><td>2016</td></tr>
<tr><td>2015</td><td>2015</td><td>2015</td><td>2015</td></tr>
<tr><td>2014</td><td colspan="10" rowspan="4">2010</td><td>2014</td><td>2014</td><td rowspan="3">2011</td><td>2014</td></tr>
<tr><td>2013</td><td>2013</td><td>2013</td><td>2013</td></tr>
<tr><td>2012</td><td>2012</td><td>2012</td><td>2012</td></tr>
<tr><td>2011</td><td>2011</td><td>2011</td><td>2011</td></tr>
<tr><td>2010</td><td>2010</td><td>2010</td><td rowspan="2">2008</td><td>2010</td></tr>
<tr><td>2009</td><td colspan="10" rowspan="3">2005</td><td>2009</td><td>2009</td><td>2009</td></tr>
<tr><td>2008</td><td>2008</td><td>2008</td><td>2008</td></tr>
<tr><td>2007</td><td>2007</td><td>2007</td><td rowspan="3">2005</td><td>2007</td></tr>
<tr><td>2006</td><td>2006</td><td>2006</td><td>2006</td></tr>
<tr><td>2005</td><td>2005</td><td>2005</td><td>2005</td></tr>
<tr><td>2004</td><td colspan="10">2000</td><td>2004</td><td>2004</td><td>2002</td><td>2004</td></tr>
</table>

これらの変数を基に、児童がその地域で意識的・無意識的に獲得しうる経験が、作文における表現にどのような影響を与えるのかを検討することができる。

なお、本データセットは 2004 年から 2023 年までのデータを纏めている点には留意しなければならない。

データの結合に際しては作文が書かれた当時から見て最新のデータを割り当てている。例えば、2004 年の作文データに対する図書館数は、2002 年の社会教育調査に基づいて結合されている。この対応を表 3 に示す。

以上のデータを用いて、分析を行っていく。

4．分析の結果

4．1．潜在意味解析の結果

事前に作文データの形態素解析を行い TF-IDF を求めた上で、潜在意味解析を行った。分析の結果を表 4 と図 1 に示す。なお、潜在意味解析には Pyhthon における Gemsim の LSI モデルを用いた。

表 4．潜在意味解析の結果

	Topic1	Topic2	Topic3
生活	0.256		
大切	0.238		0.107
私たち	0.237	0.393	-0.095
思い	0.211	0.177	0.173
できる	0.210		-0.123
考え	0.178		
思う	0.176		-0.067
日本	0.169	-0.262	-0.283
自然	0.156	0.268	
人々	0.144		
思っ	0.132		
使っ	0.130	-0.170	
水道	0.129	-0.225	
必要	0.127		
人間	0.120	0.166	
蛇口	0.119	-0.164	
祖父	0.102	-0.199	0.652
きれいな	0.101	0.101	
流れ	0.101	0.160	
使う	0.101		
井戸		-0.207	0.357
ダム		0.199	0.253
世界		-0.153	-0.184
節水		-0.148	-0.079
水道水		-0.128	-0.067
安全		-0.116	-0.080
ゴミ		0.115	
きれい		0.103	
生き物		0.102	
地球			-0.118
問題			-0.08
仕事			0.072
井戸水			0.070
建設			0.063
田んぼ			0.062
当たり前			-0.062
Silhouette Score			0.211
Calinski-Harabasz Index			12.423
Davies-Bouldin Index			2.196
Perplexity			0.968
Log-likelihood			1543.587
Average Cosine Similarity			0.999

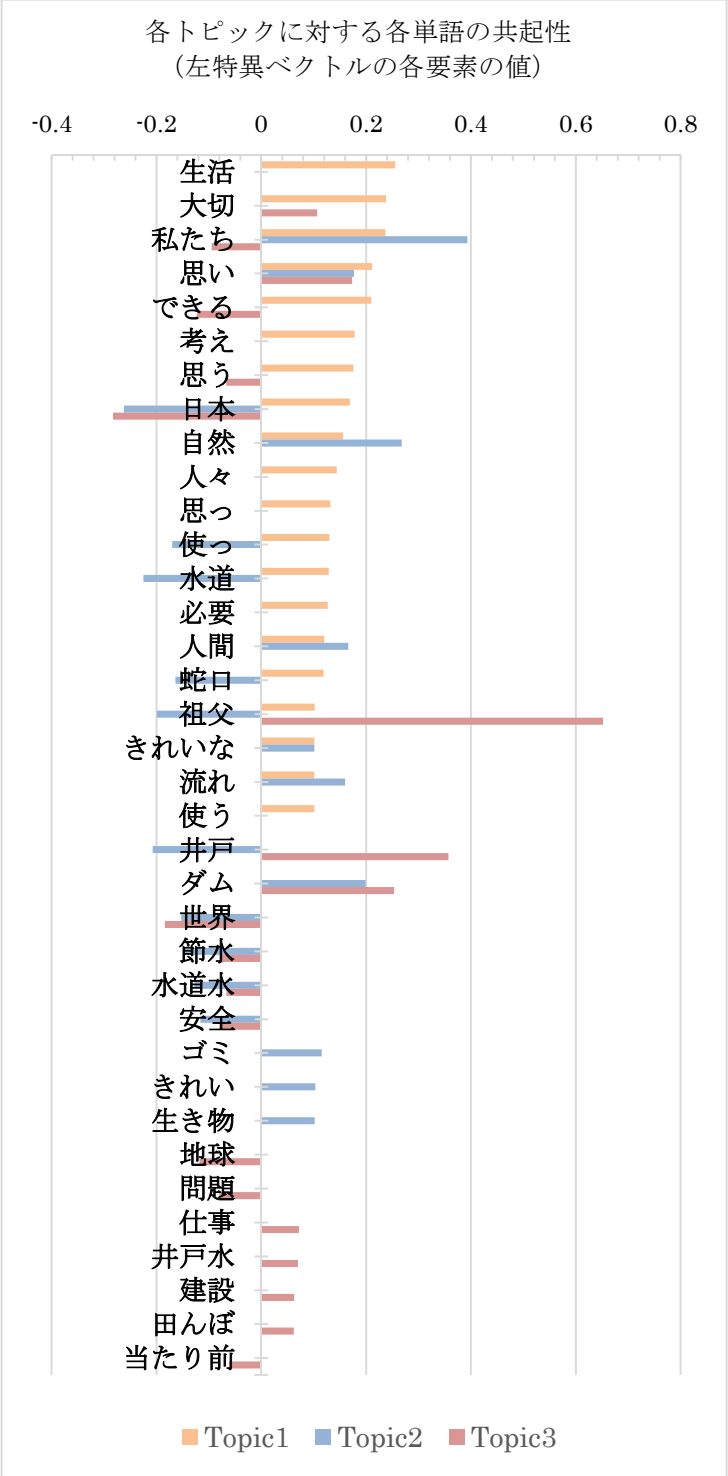

図 1. 潜在意味解析の結果

潜在意味解析の結果から、3トピックが抽出された。なお、トピック数の決定に際しては、シルエットスコア(Silhouette Score)、Calinski-Harabasz 指数、Davies-Bouldin 指数、パープレキシティ(Perplexity)、対数尤度(Log-likelihood)、平均コサイン類似度(Average Cosine Similarity)を総合的に勘案して決定した。

まず、トピック 1 は日常「生活」における水への「考え」や「思い」を全般的に抜き出したトピックと言えよう。実際に、国土交通省の HP で公開されている作文を読んでみると、「水道」や「蛇口」を「使っ」ていくなかで、水の「大切」さや「人間」にとって「必要」不可欠なものであることを再認識する作文が数多く見られる。そのため、トピック 1 を「日常生活における水トピック」と命名する。

トピック 2 は正の係数として「私たち」と「自然」や「生き物」、「きれい」な水の「流れ」を含むとともに、負の係数として「日本」や「世界」の「水道」「井戸水」「蛇口」などの水の供給システムをも含んだトピックである。なお、ここでの係数の正負はトピックを構成する単語の相対的な位置関係を現わすものである。そのため、トピック 2 は水利用を巡る自然と人工の対比を中心としたトピックと言える。トピック 2 を「水の流れ、供給システムトピック」と命名する。

トピック 3 は値が大きい「祖父」などの先人による「井戸」「井戸水」や「田んぼ」などの畑「仕事」、「ダム」を巡る経験が正の係数を示す一方で、負の係数を示すものとして生徒自身が「私たち」に「できる」ことを考え、「日本」や「世界」、さらには「地球」規模の水「問題」について考えるトピックになっている。そのため、トピック 3 を「先人の経験と、今の私たちを取り巻く水問題トピック」と命名する。

表 4 . Elastic Net 回帰の結果

		Topic1	Topic2	Topic3
1	国立私立ダミー			
2	核家族世帯数			
3	母子世帯数			
4	父子世帯数			
5	60 歳以上人口			
6	農林漁業者	0.060		-0.073
7	農林漁業雇用者	0.027		
8	労務作業者			
9	会社団体役員			
10	教員・宗教家			
11	文筆家・芸術家・芸能家			
12	1 校当たり教員数			
13	1 校当たり生徒数			
14	教育費			
15	中学校費			
16	社会教育費	0.002		
17	図書館数			
18	一級河川延長			
19	二級河川延長			
20	準用河川延長			
21	主要湖沼面積	-0.009	0.073	-0.045
22	平均気温	0.017		
23	日最高気温平均			-0.007
24	日照時間の合計			-0.005
25	降水量の合計	0.016		
26	降雪量の合計			0.023
	λ	0.205	0.076	0.076
	α	0.150	0.750	0.490

注. 空白の個所は係数が 0 であることを意味する。

潜在意味解析の結果から全 715 作文で書かれた表現を 3 つのトピックとして抽出した。次節では、このトピックに対する各作文の重みを従属変数として、これら表現を生み出す規定要因分析を行う。

4．2．Elastic Net 回帰の結果

潜在意味解析の結果を踏まえて、各トピックへの文書の重み（右特異ベクトル）を従属変数、表 1 に示した各変数を独立変数とした Elastic Net 回帰を行った。分析の結果を表 4 に示す。なお、分析には R の glmnet パッケージ(Version 4.1-8)(Friedman et al., 2023)[3]を用いた。

まず、分析に際して各トピックに対する回帰式におけるパラメータ λ と α のチューニングを行った。その結果、トピック 1 で λ=0.205; α=0.150、トピック 2 で λ=0.076; α=0.750、トピック 3 で λ=0.076; α=0.490 を得た。

まず、トピック 1 では、「農林漁業者」(β=0.060)「農林漁業雇用者」(β=0.027)が比較的高い係数を有していることが注目に値する。そもそも、トピック 1 は水に関する全般的な記述から構成される「日常生活における水トピック」であり、3 トピックの中では最も基本的かつそれ故に重要となる。そのトピックに対して、「農林漁業者」「農林漁業雇用者」が高い正の係数を有しており、本コンクールのよ

うな水に関する作文を書きあげる上では第 1 次産業の人々と陰に陽に関わることの重要性が示唆される。また、「平均気温」(β=0.017)、「降水量の合計」(β=0.016)も弱い正の係数を示している。そのため、気温や降水量の高さによって、日常的に水を意識する機会が多くなり、トピック 1 のような作文を書く傾向にあるのかもしれない。また、非常に小さい係数であるが、「社会教育費」(β=0.002)も正の値を示していることから、図書館や博物館といった地域の社会教育施設が充実していることでトピック 1 型の作文を書くことも示唆される。

トピック 2 の「水の流れ、供給システムトピック」については「主要湖沼面積」(β=0.073)のみが正の係数を示した。水の流れや生態系、そして供給システムについて書く際にはその水源を理解しなければならないため、このような結果を示しているのかもしれない。

最後に、トピック 3 については「降雪量の合計」(β=0.023)が正の効果を示す一方で、「農林漁業従事者」(β=-0.073)、「主要湖沼面積」(β=-0.045)、「日最高気温平均」(β=-0.007)、「日照時間の合計」(β=-0.005)が負の結果を示した。Topic3 は「先人の経験と、今の私たちを取り巻く水問題トピック」であるため、水を巡る問題と向き合う機会の多い地域で書かれる傾向にあるのかもしれない。その鍵となるのが「降雪量の合計」であり、毎年やってくる自然環境問題を意識するなかで、より広く「水問題」に目を向けている可能性がある。なお、「農林漁業従事者」や「主要湖沼面積」が負の係数を示しているため、一見すると都市出身者がトピック 3 型の作文を書くと考えられる。しかし、「教育費」の係数が 0 と推定された点や、「教育費」と「農林漁業従事者」(r = 0.113)、「主要湖沼面積」(r = -0.042)の間でそれぞれ無相関であったため、必ずしも都市出身者がトピック 3 型の作文を書く傾向にあるとは言えない。なお、トピック 3 の単語構成で最も大きな要因である「祖父」からして、60 歳以上人口が効きそうであるが、係数が 0 と推定された点は意外な結果であった。

5．結論と考察：言語による表現の規定要因として人的環境の重要性

本稿では中学生が言語によって繰り出す表現の規定要因を、地域の社会経済環境や自然環境、人的環境に着目して明らかにすることを目的として分析を展開してきた。

以上の分析から本稿の知見は以下のように整理できる。

第一に、本研究対象である「水の作文」における表現は、生徒が日々経験する地域の自然環境に基づいて書かれていることが示唆された。ここでもう一度、渡邉雅子の議論を振り返れば、日本の作文教育の特徴は児童の体験を基にした共感型の作文を書くことにあった(渡邉, 2023)[14]。この指摘を踏まえれば、Elastic Net 回帰において、各トピックに対して自然環境変数が影響を示していたことも納得ができよう。この渡邉の指摘は我々の経験を顧みても納得がいくものであるが、これまで児童生徒の作文データを基に統計的に検討したものではなかった。そのため、本研究の知見の一つはこの渡邉の指摘を統計的に実証したことにある。

重要なのは第二の知見であり、生徒の作文における表現が地域の人的環境によってもたらされる可能性があり、かつその影響の大きさは社会経済環境や自然環境よりも大きいことを示唆した点にある。地域の財政規模や自然環境は個人にとっては所与の条件であり、これらを個人の力で変えることは困難である。そのため、これらの条件が作文のような個人の成果物に影響を与える場合、場合よっては「格差」を孕んでいるとも解釈できる。一方で、人的環境はより個人の主体性を発揮しやすく、個人が様々な人と関わろうとすることで成果物に対して正の影響を与えるのであれば、社会経済環境などの所与の条件を乗り越える変数として期待される。このような、人との関係性による効用は政治学や社会学で社会関係資本(Social Capital)と呼ばれ、特に Putnam(2000=2006)[9]等が著名な研究として知られる。「言語力」との関係についても、志水ほか(2014)[10]や山田(2014)[12]によって国語のテスト得点との関係で、その効果が提示されてきた。しかし表現というより抽象的な次元においても人的環境(社会関係資本)の効果が示唆された点は、潜在意味解析と Elastic Net 回帰を用いて作文データへの多変量解析を実現した本研究ならではの示唆であり、本研究の最大の知見である。

ただし、今回の研究ではあくまで個人データに記載された出身地と基幹統計等を結び付けたデータに基づくため、個人が地域の社会経済環境や自然環境、人的環境に接触しようとする主体性までは考慮できていな

い。この点は今後の課題としたい。

注

1) ここでいう「言語力」は、本文で言及した学習指導要領に規定される「言語能力」や PISA 調査の「読解リテラシー」などを含んだ包括的な概念として記している。
2) 作文が書かれた当時は存在していたが、2024 年時点までに閉校した学校や、自治体の合併などで設置主体が町立から市立になったため厳密には当時のままの学校が存続していないケースなどが 149 名・106 校見られた。これらの学校については e-Stat の市町村別データの出力に際して、絞り込み段階で「表示データ」を「現在の市区町村」で出力している関係上、当時の学校の所在地を現在の行政区分に基づく所在地で整理し、結合している。なお、行政区分は現在の区分であるが、データは表 3 の通りである。

参考文献

[1]荒川清晟・野寄修平：“大都市から地方への移住と社会経済的要因の関連—Elastic Net 回帰を用いたポアソン重力モデルによる分析—”, 社会情報学, 11, 3, pp.19-33(2023).

[2]Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. : “Indexing by latent semantic analysis”, *Journal of the American Society for Information Science*, 41(6), pp.391-407(1990).

[3]Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., Qian, J & Yang, J., “Package ‘glmnet’ ”, (2023), https://cran.r-project.org/web/packages/glmnet/index.html(最終確認日：2024/08/05).

[4]Hoff, E. : “How social contexts support and shape language development”, *Developmental Review*, 26(1), pp.55-88(2006).

[5]川野秀一・松井秀俊・廣瀬慧：“スパース推定法による統計モデリング”,pp.29-46, 共立出版(2018).

[6]国土交通省, https://www.mlit.go.jp/mizukokudo/mizsei/tochimizushigen_mizsei_tk1_000010.html(最終確認日：2024/05/25)

[7]文部科学省：“中学校学習指導要領(平成 29 年告示)”(2017)

[8]文部科学省・国立教育政策研究所：“OECD 生徒の学習到達度調査 PISA2022 のポイント”(2023) https://www.nier.go.jp/kokusai/pisa/pdf/2022/01_point_2.pdf(最終確認日: 2024/07/31).

[9]Putnam, R.D. : “*Bowling Alone: The Collapse and Revival of American Community*” Simon and Schuster,(2000) (=柴内康文[訳]“孤独なボウリング—米国コミュニティの崩壊と再生—”, 柏書房, (2006))

[10]志水宏吉・中村瑛仁・知念渉：“調査報告「学力格差」の実態”, 岩波ブックレット(2014).

[11]鳶島修治：“読解リテラシーの社会経済的格差—PISA2009 のデータを用いた分析—”, 教育社会学研究, 98, pp.219-239(2016).

[12]山田哲也：“社会経済的背景と子どもの学力(1) 家庭の社会経済的背景による学力格差：教科別・問題別・学校段階別の分析”, 平成 25 年度 学力調査を活用した専門的な課題分析に関する調査研究, pp.57-70(2014).

[13]渡邉雅子：“作文指導に見る個性と創造力のパラドックス—日米初等教育比較から—”, 教育社会学研究,69, pp.23-42(2001).

[14]渡邉雅子：“日・米・仏の国語教育を読み解く：『読み書き』の歴史社会学的考察”, 日本研究, 35, pp.573-619(2007).

[15]渡邉雅子：“「論理的思考」の文化的基盤—4 つの思考表現スタイル”, 岩波書店(2023).

[16]Zou, H.& Hastie, T : “Regularization and variable selection via the Elastic Net”, *Journal of the Royal Statistical Society Series B*, 67(2), pp.301-320(2005).