新聞投稿文を特徴づける因子の抽出

生田　和重

徳島文理大学　文学部

1　はじめに

学生の日本語文章力を高めるために、筆者が担当する授業において「分かりやすい実用文の書き方」を伝授している（生田 2002)。この授業で、学生はオンライン教材を参考にしつつ作文の基本を身に付ける。その学習効果を自ら把握するために、「日本語文章能力検定」に挑戦する。さらに最終課題として、朝日新聞の声欄に投稿する。

この学習を 2001 年度から継続して実施し、学生が作成した投稿文を電子データとして蓄積している。そして、この投稿文データを定量的に分析し、その結果をもとに設定したチェック項目を投稿前に確認させたところ 、声欄への掲載率が大幅に改善された（生田ほか 2005a、生田ほか 2005b)。

今後とも、学生の立場に立った改善を加えながら、この学習を継続していきたいと考えている。生田 (2007) では、投稿文のテーマ選定で参考にしてもらうために、「声欄への投稿文」で使われている語彙を統計的に分析し、その特徴を把握した。その際、先行研究（宋 2003b）の分析結果（朝日新聞の「オピニオン」で使われている語彙の出現頻度データ）との比較を試みた。その結果、「声欄への投稿文」では「家族」や「学校」に関連する形態素の出現頻度が顕著に高いことが分かった。しかし、両者（生田 2007 と宋 2003b）の分析方法には違いが見受けられ、厳密な比較になっていないという危惧が残った。そこで本論文では、「声欄への投稿文（読者からの投書)」と「私の視点（専門家や有識者による意見文)」の語彙を対象に、独自の手順で語彙頻度分析と因子分析を実施する。そして、両分析結果を比較することにより、「声欄への投稿文」の特徴を明らかにしたい。

2　分析方法

2.1　分析対象データと比較データ

学生の投稿文は、E メールで朝日新聞編集局「声」係（dai-koe@asahi.com）へ送信される。そして、編集者の眼鏡にかなった投稿文は朝日新聞大阪版の朝刊に掲載される。そこで、分析対象データを朝日新聞大阪版に掲載された「声欄への投稿文（読者からの投書)」とした。また、比較データとして朝日新聞大阪版に掲載された「私の視点（専門家や有識者による意見文)」を準備した。なお分析の際には、筆者と朝日新聞社との間で利用許諾契約を結んでいる「朝日新聞記事データ集 2002（朝日新聞社 2003)」を活用した。

2.2　語彙頻度分析の手順

最初に、適切なキーワードを指定して「朝日新聞記事データ集 2002（朝日新聞社 2003)」から分析対象データ（声欄への投稿文）と比較データ（私の視点）を抽出した。分析対象データについては、月別にファイル名を付けて、テキスト形式で保存した。比較データについては、1 年間のデータをまとめてテキスト形

式で保存した。その後、各テキストファイルに語彙頻度分析（藤井ほか 2005，林 2002) を施した。その手順は以下の通りである。

まず、「形態素解析システム茶筌（WinCha，松本ほか 2000)」にテキストファイルの中身をコピー＆ペーストする。「基本形、品詞」のみにチェックマークを付けて全文解析を実施し、解析結果を保存する。その解析結果を表計算ソフトで読み込む。この解析結果データは、「形態素」列と「品詞」列から構成される。つぎに、形態素が属する投稿文（または私の視点）を区別するために、各々に ID（番号）を付与する。形態素解析の結果データには、各文章の終わりに記号「EOS(End of Sentence)」が追加される。この記号を判定条件として、1 から順に ID を付けた。また、得られた形態素の中には、その文章の特徴を表すキーワードとしては不適切なものも含まれる。そこで、藤井ほか (2005) と林 (2002) を参考にして、適切な品詞のみを抽出する。抽出する品詞は、「形容詞－自立」、「形容詞－接尾」、「形容詞－非自立」、「名詞－サ変接続」、「名詞－一般」「名詞－形容動詞語幹」、「名詞－固有名詞語幹」である。具体的には、解析結果データに「品詞条件（採択／非採択を判定する条件式)」の列を追加し、「採択」のみを残した。

このようにして抽出された各形態素に、分類語彙データベース（国立国語研究所 2004）を活用して、分類コードグループ（表 1 ）を自動で付与した。その際、分類語彙データベースに含まれていない（出現頻度が稀な）形態素については分析対象から除外した。今回の分析で対象から除外された形態素の割合は全体の 7 %程度であった。

表 1　分類コードグループと項目との対応

分類コードグループ（分類コード）	概要
G 1 （1.10～1.15）	抽象的関係（体）
G 2 （1.16, 1.17）	位置、期間、場所（体）
G 3 （1.18, 1.19）	形、量（体）
G 4 （1.20～1.22）	人間活動の主体（私的、体）
G 5 （1.23～1.25）	人間活動の主体（公的、体）
G 6 （1.26～1.28）	社会的場所、機関、団体（体）
G 7 （1.30～1.33）	精神および行為（私的、体）
G 8 （1.34～1.38）	精神および行為（公的、体）
G 9 （1.40～1.47）	生産物および用具（体）
G 1 0 （1.50～1.56）	自然物および自然現象（体）
G 1 1 （1.57, 1.58）	体、生命、健康（体）
G 1 2 （3.10～3.17, 3.19）	抽象的関係（相）
G 1 3 （3.30, 3.33, 3.34, 3.36, 3.37）	精神および行為（相）
G 1 4 （3.50～3.52, 3.55, 3.57, 3.58）	自然現象（相）

最後に、分類コードグループと ID とのクロス集計を行い、出現頻度を算出した。なお、分類コードグループ毎の出現率は、以下の式で求めた。ここで総出現頻度は、各分類コードグループの出現頻度の総和である。

$$\text{分類コードグループ毎の出現率（\%）} = \frac{\text{分類コードグループ毎の出現頻度}}{\text{総出現頻度}} \times 100$$

2.3　因子分析の手順

前節の手順で作成した分類コードグループと ID とのクロス集計データに対して、SPSS12.0J (SPSS Inc. 2003) を使用して因子分析（石村 2005）を施した。因子抽出法は、「最尤法」を用いた。まずは因子数と回転方法を指定せずに分析し、スクリープロット図をもとに因子数を決定した。その際には、「KMO および Bartlett の検定（石村 2005)」によって、今回のデータを対象にして因子分析を行うことの妥当性を判定した。その後、決定した因子数を入力し、回転方法としてプロマックス回転（斜交回転）を選択して、正式な分析を実施した。なお、「声欄への投稿文」については、月別データを対象にして別々に分析すると、抽出される因子にばらつきが生じることが懸念された。そこで、6 ヶ月分をまとめたデータを作成し直して、それを対象に因子分析を実施した。

3　分析結果と考察

図１と図２に、語彙頻度分析によって得られた「分類コードグループ毎の出現率」を示した。ここで、各々の分析対象データは、2002 年１月掲載の「声欄への投稿文」と 2002 年掲載の「私の視点」である。両図を対比すると、いずれの場合にもＧ７「(私的な）精神および行為」とＧ１「抽象的な関係」の出現率が顕著に高いことが分かる。そして、Ｇ４、Ｇ９、Ｇ６、Ｇ８の出現率については両者に明らかな違いがあることに気がつく。すなわち、Ｇ４とＧ９については、「声欄への投稿文」（図１）の出現率が「私の視点」（図２）のそれの２倍以上である。一方、Ｇ６とＧ８については、「私の視点」の出現率が「声欄への投稿文」のそれよりも 50 %以上大きい。ここで、Ｇ４は「人間活動の主体（家族や仲間など)」を、Ｇ９は「(日常生活で使用する）生産物や用具など」を包括する分類コードグループである。さらに、Ｇ６は「社会的場所、機関、団体」を、Ｇ８は「(社会の一員としての公的な）精神および行為」を含む。この結果から、「声欄への投稿文」では「個人的な体験」、「家族との思い出」、「日常の悩みや楽しみ」などの私的なテーマが大勢を占めていると推察できる。一方、「私の視点」は専門家や有識者による「社会問題に対する意見文」という位置づけであるので、Ｇ８やＧ６の出現率が目立つのは当然かもしれない。

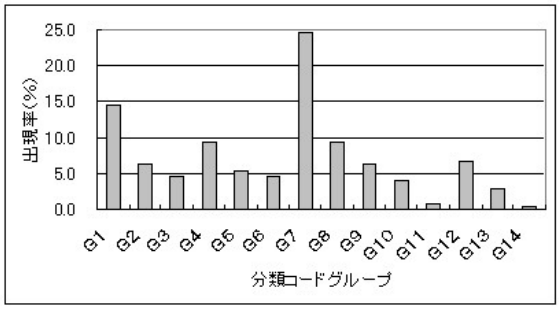

図1 「声欄への投稿文」の分類コードグループ毎の出現率
(2002年1月、総出現頻度＝10745)

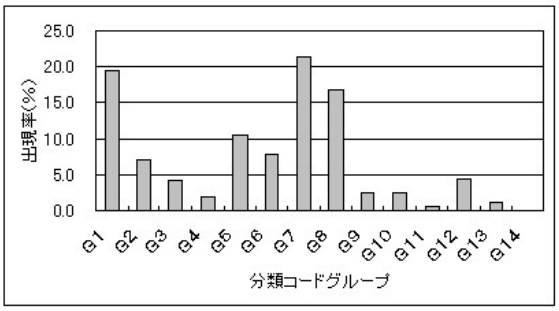

図2 「私の視点」の分類コードグループ毎の出現率
(2002年、総出現頻度＝9803)

つぎに、「最尤法」で因子分析を施した結果を表２と表３に示した。なお、因子数と回転方法を指定せずに事前分析を行い、得られたスクリープロットをもとに因子数を「２」と決定した。また、回転方法として「プロマックス回転」を用いた。表２の第１因子が分類コードグループ「Ｇ１０，Ｇ９，Ｇ１４，Ｇ２」に影響を与えることから、この因子を「自然物や生産物」と命名したい。また、第２因子が「Ｇ８，Ｇ１，Ｇ５」に影響を及ぼすことから、この因子を「社会における公的な活動」と名づけても良いであろう。ここで、「Ｇ７」と「Ｇ４」が、各々、第１因子と第２因子に対して顕著な負の因子負荷を示していることの意味は定かではない。しかし、「私的な精神や行為」が含まれていない単なる「自然、生物、生産物」の紹介の場合には正の因子負荷となり、逆にそれが含まれる場合には負の因子負荷となるとも考えられる。そこで、「声欄への投稿文」の第１因子を「自然物や生産物と私との係わり」と改名する。同様に、「私的な人間活動の主体」が表にでない「社会における公的な活動」では正の因子負荷となり、逆にそれが含まれる場合には負の因子負荷となると考えられる。そこで、「声欄への投稿文」の第２因子を「社会生活と私との係わり」と改称する。結論として、「自然物や生産物と私との係わり」と「社会生活と私との係わり」という２つの因子で「声欄への投稿文」を特徴づけられると考える。表３の第１因子が分類コードグループ「Ｇ１，Ｇ３，Ｇ６，Ｇ８，Ｇ２，Ｇ１２，Ｇ９，Ｇ１０」に影響を与えることから、この因子を「社会生活や社会

表２　「声への投稿文」の因子パターン

分類コードグループ	第１因子	第２因子
Ｇ１０	0.634	0.086
Ｇ９	0.531	0.063
Ｇ１４	0.400	0.002
Ｇ２	0.349	0.224
Ｇ１２	0.183	0.035
Ｇ３	0.161	0.014
Ｇ１１	0.070	-0.038
Ｇ８	-0.107	0.653
Ｇ１	0.031	0.533
Ｇ５	-0.217	0.449
Ｇ６	-0.247	0.288
Ｇ７	-0.431	0.105
Ｇ１３	-0.131	-0.182
Ｇ４	-0.251	-0.541

表３　「私の視点」の因子パターン

分類コードグループ	第１因子	第２因子
Ｇ１	0.888	0.042
Ｇ３	0.853	-0.262
Ｇ６	0.778	-0.056
Ｇ８	0.703	-0.032
Ｇ２	0.551	0.134
Ｇ１２	0.532	0.342
Ｇ９	0.509	-0.115
Ｇ１０	0.323	0.008
Ｇ５	0.280	0.261
Ｇ７	0.007	0.910
Ｇ１１	-0.156	0.576
Ｇ１３	0.187	0.318
Ｇ４	-0.101	0.264
Ｇ１４	-0.336	0.026

問題（への提言）」と命名したい。また、第２因子が「Ｇ７，Ｇ１１，Ｇ１３」に影響を及ぼすことから、この因子を「生命や医療（への提言）」と名づけても良いであろう。

4　おわりに

朝日新聞の「声欄への投稿文」は比較的に身近なテーマについて自らの経験や考えをまとめた文章であると総括できる。投稿するという行為を通して、その作者は日常のストレスを発散したり、自分の考えや行動の妥当性を確認したりしているのであろう。一方「私の視点」は公的な印象が強く、その主体は社会生活や社会問題への提言であると判断できる。このように、語彙頻度分析や因子分析によって原文の大まかな特徴を把握できることは意味深いと考える。今後は、この分析結果を「分かりやすい実用文の書き方」の学習において有効活用していきたい。さらに、「声欄への投稿文」や「私の視点」の二文間の連接関係やストーリー展開について分析し、その特徴を総合的に把握する予定である。なお、本研究の一部は文部科学省科学研究補助金基盤研究（B)（課題番号 19300292、代表　石岡恒憲）の援助を受けた。

参考文献

朝日新聞社 (2003)　朝日新聞記事データ集 2002　学術・研究用．日外アソシエーツ株式会社，東京
藤井美和ほか (2005)　福祉・心理・看護のテキストマイニング入門．中央法規出版，東京
林俊克 (2002)　Excel で学ぶテキストマイニング入門．オーム社，東京
生田和重 (2002)　学内 LAN を活用した文科系学生に対する授業実施例．教育システム情報学会誌, 19(1), pp.28-32
生田和重ほか (2005a)　文科系学生が作成した投稿文の統計的な分析．日本教育工学会論文誌，29(1)，pp.35-42
生田和重ほか (2005b)　文科系学生が作成した投稿文の統計的な分析とその結果を活用した学習事例．日本行動計量学会第 33 回大会発表論文抄録集，pp.382-385
生田和重 (2007)　語彙分析による新聞投稿文の特徴把握．言語処理学会第 13 回年次大会発表論文集，pp.332-335
石村貞夫 (2005)　SPSS による多変量データ解析の手順 [第３版]．　東京図書，東京
国立国語研究所 (2004)　分類語彙表－増補改訂版－　データベース．国立国語研究所，東京
松本裕治ほか (2000)　形態素解析システム茶筌．http://chasen.aist-nara.ac.jp/hiki/ChaSen/
宋正植 (2003b)　『朝日新聞』の「オピニオン」の比較語彙研究　－ 1946 年と 2000 年の語彙比較を通して．名古屋大学大学院国際言語文化研究科研究誌『ことばの科学』 第 16 号，pp.27-54
SPSS Inc. (2003) SPSS 12.0J Brief Guide. SPSS Inc. ，東京