語彙分析による新聞投稿文の特徴把握

生田　和重

徳島文理大学　文学部

１　はじめに

近年、大学生の日本語文章力が著しく低下しており、日常のレポートにおける論理構成の不可解さは私たち教師を悩ましている。また、学生が就職活動で自己 PR や志望動機を作成する際に、「文章の勉強をしておけば良かった」と悔いているのを頻繁に見かける。このような状況に対応するために、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（生田 2002, 生田 2003）」を伝授している。この授業で学生は、オンライン教材（生田 2002）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（日本語文章能力検定協会 2006)）」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（生田 2002)。

この学習を 2001 年度から３年間継続して実施し、学生が作成した投稿文を電子データとして蓄積した。そして、このデータと声欄に掲載された投稿文のデータとを、修辞項目について、定量的に比較した。その結果、最大文字数と語彙の多様性ついて両者の間で有意な差があることが判明した（生田ほか 2005a)。この結果をもとに、最大文字数を７０字以下にする、多様な語彙を用いるために投稿文のテーマについて十分に調査する、などのチェック項目を設定した。このチェック項目を全て満足しているか否かを、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 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。まず本研究では、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、その特徴を把握する。

２　先行研究

新聞記事の語彙を分析した先行論文として、柏野ほか(2005)、宋(2003a) 、宋(2003ｂ)を挙げることができる。

柏野ほか(2005)は、「外来語を対象に、新聞記事データベースを用いて、語の使用推移、すなわち出現率の推移に着目して語の使用状況をとらえる研究」である。この論文で彼らは、「出現率＝使用度数÷収録文字数」という式を用いて、使用度数を正規化している。そして、その理由について、「本稿では収録単語数の代わりに、より算出の容易な収録文字数を用いて正規化することとした」と説明している。今回の論文では、第３章に示すように、総単語数（単語の総出現回数）を用いて、より厳密な正規化を行いたい。

また宋(2003a) では、「朝日新聞の天声人語を対象にした 1946 年と 2000 年の語彙比較」を実施している。さらに宋(2003ｂ)では、「朝日新聞のオピニオンを対象にした 1946 年と 2000 年の語彙比較」を実施している。朝日新聞のオピニオン欄には、「オピニオン１：視点」と「オピニオン２：声（投稿文)」が掲載される（朝日新聞社 2003)。その内容から、宋(2003ｂ)は「オピニオン１：視点」を対象とした分析であると推測される。すなわち両論文は、朝日新聞の編集者や諸分野の見識者が書いた文章に含まれる語彙を分析して、時代背景を把握しようと試みたものであろう。また、これらの論文では、「文節を第一次の調査単位とし、そこから１つの自立語と、

いくつかの付属語を析出」している。今回の論文では、藤井ほか(2005)と林(2002)を参考にして、「対象とする語彙」を抽出する。すなわち、まず「形態素解析システム茶筌（WinCha，松本ほか 2000）」で投稿文を形態素に分解する。その後、対象とする品詞を指定して「対象とする形態素の集合（語彙）」を抽出する。

３　データ分析

３．１　分析対象

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、分析対象を「朝日新聞大阪版の声欄に掲載された投稿文」とした。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。

３．２　分析方法

最初に、キーワード「投書＊声＊大阪」で検索して、「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を抽出した。この分析対象データを、月別にファイル名を付けて、テキスト形式で保存し、各テキストファイルに形態素解析を施した（藤井ほか 2005，林 2002)。この形態素解析には、「形態素解析システム茶筌（WinCha，松本ほか 2000）」を使用した。

得られた形態素の中には、投稿文の特徴を表すキーワードとして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、分析に使えそうな品詞を抽出した。この際、抽出した品詞は、「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。そして、形態素毎の「出現頻度（総計）」を求めた。なお本論文では、「出現頻度（総計）」が３以上の形態素のみを分析の対象とした。さらに、得られた各形態素に、国立国語研究所 1964，国立国語研究所 2004a，国立国語研究所 2004b を参考にして分類コードを付与した。

上記の処理を施した後、分類コードで並べ替え、分類コード毎にデータを集計して出現頻度を算出し、以下の式で正規化した。ここで総出現頻度は、各分類コードの出現頻度の総和である。

分類コード毎の出現率＝分類コード毎の出現頻度／総出現頻度

４　分析結果と考察

図１に、2002 年 1 月の「分類コード別の出現率分布」を示した。体の類（名詞）で、最も頻繁に出現する分類コードは「1.30（心・学習・知識・思考・原理・規則・主義）」であり、出現率は 10%を超えている。「1.21（家族）」と「1.20（われ・なれ・かれ・だれ・人間・神仏）」がそれに続き、その出現率は 5～7%である。それらに続く、「1.26（社会的場所・資産）」、「1.31（言動・語・発言・沈黙・話・読み書き）」、「1.33（文化・歴史・風俗・人生・禍福・労働・生活・学事・祭儀・遊楽）」、「1.24（成員・職）」の出現率は各々3～5%である。これら上位７つの出現率を合わせると 40%超となる。

また相の類（形容詞）では、「3.19（長さ・大きさ・速さ・重さ・量・限度）、「3.30（意識・感覚・驚き・楽・苦・好き・嫌い）」、「3.13（繁簡・普通・非凡・良・不良）、「3.12（在不在・必然性・可能性）」の出現率が目立っている。ただし、これらの出現率は各々3%未満である。この分

析結果から、「家族に対する思い」、「家族との思い出」、「個人的な体験」、「社会環境への提言や苦情」、「職場での出来事」、「日常の悩みや楽しみ」などのテーマを思い浮かべることができる。これらのテーマで投稿することにより、日常のストレスを発散させたり、自分の考えの妥当性を確認したりしているのであろう。これは、いわゆるカタルシス（感情表出による精神浄化作用）につながる行為であると推察する。また、これらのテーマであれば、事前に調査することなく、独自の経験と奇抜な発想のみを材料に投稿文を作成できる。その気軽さも、この結果の一因であろう。

つぎに、出現率が高い形態素について言及する。英オックスフォード大学出版局の調査（朝日新聞社 2006)によれば、「最も頻繁に使われる英単語（名詞)」は

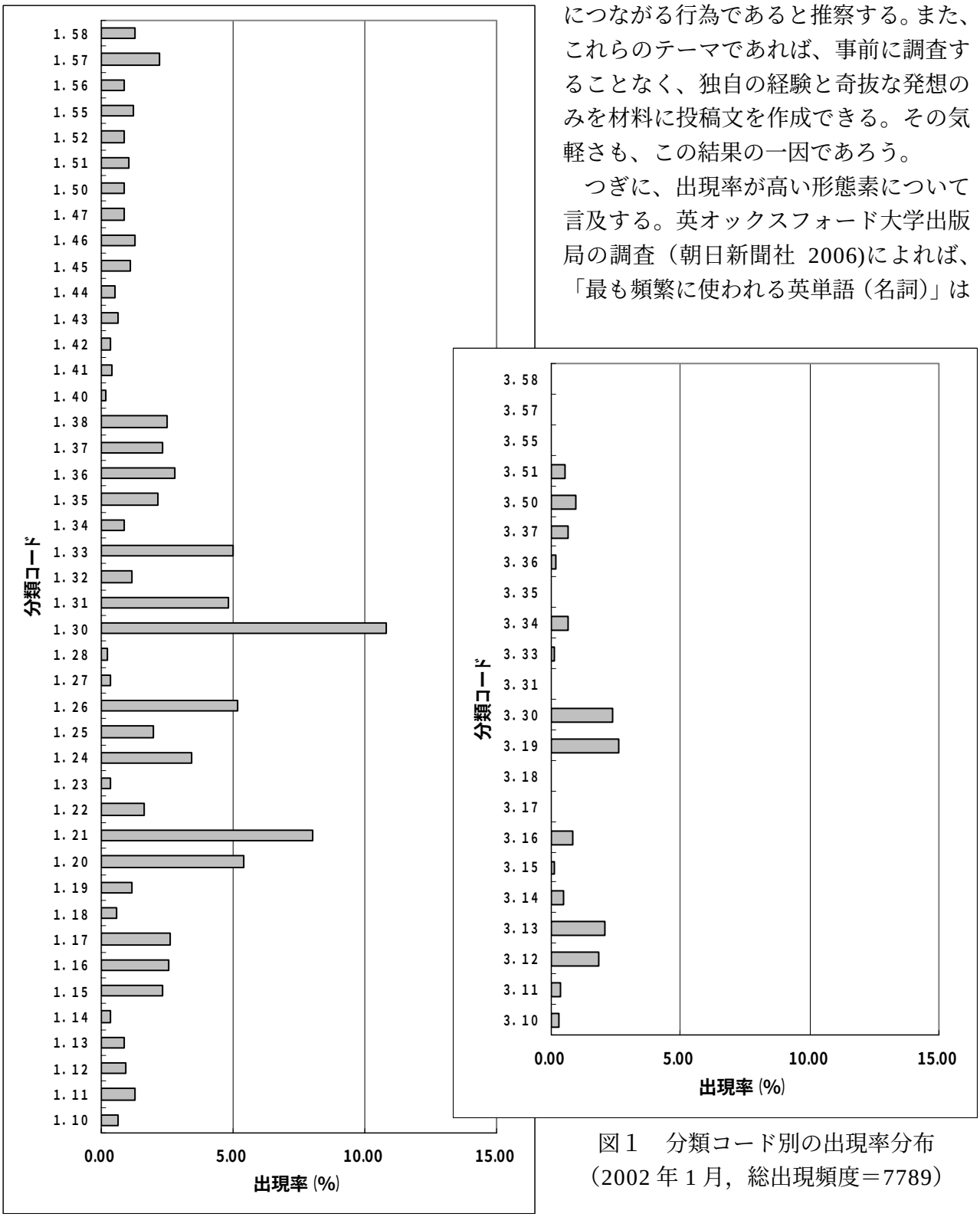

図１　分類コード別の出現率分布
（2002 年 1 月，総出現頻度＝7789）

「TIME（時間）」であるらしい。以下、「PERSON（人）」、「YEAR（年）」、「WAY（道、方向）」、「DAY（日）」が続く。同出版局は、この結果について「われわれは時間に支配されているようだ」とコメントしている。この例を参考にすれば、「出現率が高い形態素」を把握することによって、投稿文の大まかな傾向をつかむことができると考える。

月別の「出現率が高い形態素ベスト 30」では、「家族」や「学校」に関連する形態素が上位を占めている。これは、日常生活に密着した投稿文が多いことを裏付けていると考える。また、「ない」、「よい（いい）」、「多い」、「ほしい」という形容詞が 30 位以内に含まれていることは興味深い。これらの形容詞は、各々「在不在・可能性」、「適・不適」、「量」、「気持ち（要望）」を表す。したがって、この結果から、評価や要望を含む投稿文が多いと推察できる。

５　おわりに

新聞記事データベースを活用して、投稿文で使われている語彙を統計的に分析した。その結果、日常生活に密着したテーマに関する投稿文が多いことが分かった。また、その内容には評価や要望が含まれていると推察できる。

今後は、今回の分析結果を「分かりやすい実用文の書き方」の学習で活用していく。また、二文間の連接関係や文章構成について分析して、新聞投稿文の特徴を総合的に把握したい。

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

朝日新聞社(2006)　英単語の中で最も使う名詞．朝日新聞 2006 年 6 月 24 日付記事

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)：28-32

生田和重(2003)　インターネット的文章力強化法．徳島文理大学研究紀要，第 65 号：17-24

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)：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：382-385

柏野和佳子ほか(2005)　新聞記事データベースを利用した外来語の出現率の推移調査．自然言語処理，**12**(4)：97-116

国立国語研究所(1964)　分類語彙表．秀英出版，東京

国立国語研究所(2004a)　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

国立国語研究所(2004b)　全文検索システム『ひまわり』．http://www.kokken.go.jp/lrc/index.php

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/

日本語文章能力検定協会(2006)　日本語文章能力検定過去問題集．㈱オーク，京都

宋正植(2003a)　「天声人語」の比較語彙研究（その１）－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『言葉と文化』 第 4 号：179-198

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：27-54