コーパスに基づく多様なジャンルの文体比較 －短単位情報に着目して－

小磯花絵　　小木曽智信　　小椋秀樹　　宮内佐夜香
国立国語研究所研究開発部門

1　はじめに

本研究では，新聞や書籍，web 上の文章，話し言葉など，様々なジャンル，媒体におけるテキストの文体の差や類似性を，短単位（単語）から得られる情報の比較を通して探る．小磯ほか (2008a,2008b) では，同じく短単位情報に着目して白書，新聞，文学の比較を行った．今回は，改まり度のより低い，あるいは話し言葉的なデータを加えることで，多様なジャンルの文体の差を多角的に捉えることを試みる．

2　方法

2.1　分析データ

分析データとして，現在構築中の『現代日本語書き言葉均衡コーパス』（BCCWJ）と，2005 年に一般公開された『日本語話し言葉コーパス』（CSJ）を用いた．BCCWJ から，(1) 行政白書，(2) 新聞[*1]，(3) 小説，(4)WEB データ（電子掲示板サイト「Yahoo!知恵袋」），(5) 国会会議録[*2]を，CSJ から，(6) 学会講演，(7) 模擬講演（主に個人的内容に関する一般の人によるスピーチ）を選択した．この分類を本研究では便宜的に「ジャンル」と呼ぶ．

各ジャンルからそれぞれ 150 サンプルを抽出して分析に用いた．CSJ については，一つの講演を１サンプルとした．また BCCWJ については，節や章などの文章のまとまりを一つのサンプルと定める可変長データセットからサンプルを抽出した．ジャンル (4) は１サンプルが質問と回答（ベストアンサー 1 件）の対から構成される．また (5) は一つの会議全体が１サンプルとなる．

2.2　指標

本研究では「語」の品詞と語種に着目して各ジャンルの比較を行う．著者等は短単位と長単位と呼ばれる２種類の言語単位を設計し，CSJ，BCCWJ への情報付与を行ってきた．「防災基本計画として」を「防災／基本／計画／と／し／て」と細かく分割するのが短単位，「防災基本計画／として／」のように，複合名詞や複合辞を一つの単位とみなすのが長単位である（小椋ほか 2008a）．本研究ではこのうち短単位を用いて分析を行った．

短単位は形態素解析用電子化辞書 UniDic で採用されている言語単位である．UniDic には品詞情報に加えて語種の情報も付与されており，全ての見出し語に対し語種情報を出力することができる（小椋ほか 2008b）．

1050 のサンプルのうち 387 サンプルは，UniDic を用いて MeCab で自動解析したもの（解析精度 98～99.5%）を，残りは人手修正を施したものを利用した[*3]．

個々の品詞率，語種率は，サンプル毎の延べ語数に対する各品詞・語種の延べ語数の割合として求めた．ただし感動詞と句読点は集計の対象としなかった．CSJ には言い淀みに伴うフィラーが頻出しており，それらが感動詞として分類されていること，また CSJ には句読点が存在しないことがその理由である．また語種については更に，助詞，助動詞，固有名詞，記号を除外した上で比率を算出した．なお，分析対象とするサンプルの長さに幅があるため（WEB：平均 240 語～国会：平均 5.5 万語），異なり語数で見た場合の品詞率，語種率は求めず，延べ語数のみを分析対象とした．

3　ジャンル毎の品詞・語種の比較

図１に個々の指標の基礎統計量をジャンル毎に示す．本稿では紙面の都合上，小磯ほか (2008a,2008b) や小椋 (2005) で特にジャンルとの関係の深かった，漢語率，和語率，名詞（普通名詞）率，接続詞率，形容詞率，機能語（助詞・助動詞）に限定して示す．

■和語率・漢語率・名詞率・機能語率：　小磯ほか (2008a,2008b) では，同種の指標を用いて白書，新聞，文学[*4]の比較を行い，漢語率，名詞率については白書，新聞，小説の順に少なくなる傾向が，和語率，機能語率については同順に多くなる傾向が見られることを示し，このことに次の三つの要因が関係している可能性があることを指摘した．

一つは，白書には「核燃料加工施設」のような複合名詞が多く，文学には少ないという点である[*5]．複合名詞は特に専門性の高い語に多いことから，この傾向は専門

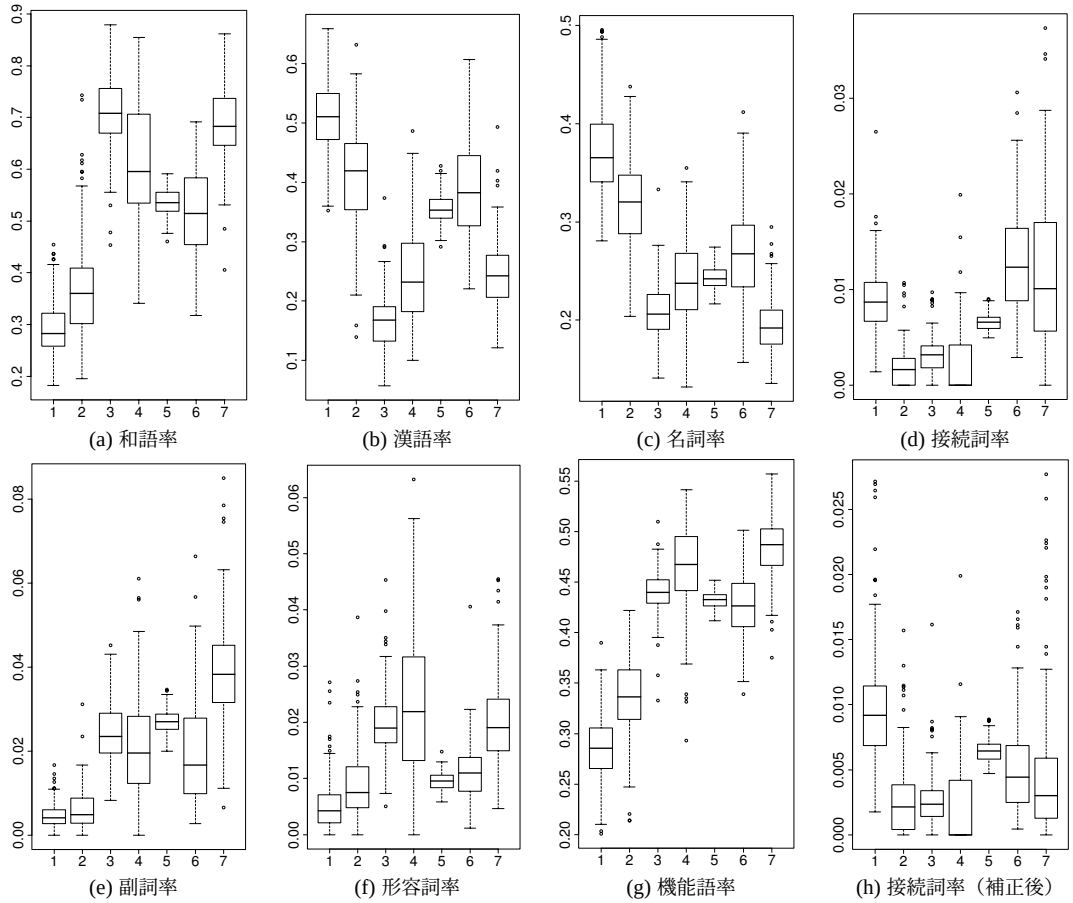
[*1] 連載小説，寄稿，図表等が紙面の多くを占める記事を除いた．

[*2] 1980～2005 年の会議を対象とした．原稿を単純に読み上げる会議をできるため避けるため，原稿読み上げ率の高いとされる本会議，短い会議・委員会 (500 文以下) を除外してサンプルを選定した．現時点で BCCWJ に含まれる国会会議録ではサンプル数が足りないため，適宜データを追加した．追加サンプルの仕様は BCCWJ に含まれるものと同じである．

[*3] CSJ にも短単位情報は付与されているが，若干基準が異なる．また語種情報はない．本研究で用いたデータは，最新の基準に合わせて人手修正し，語種情報を追加したものである．

[*4] 本研究とは異なり，小説以外に若干の書評等も含む．

[*5] 複合名詞が多いと，短単位で品詞率を算出した場合に名詞率は高くなる．また複合名詞を構成する語の大半は漢語であるため，漢語率も高くなる．一方，機能語率は概ね内容語率の逆数であり，内容語の増減は名詞率に強く依存することから，機能語率は名詞率と逆の傾向となる．



(a) 和語率　(b) 漢語率　(c) 名詞率　(d) 接続詞率

(e) 副詞率　(f) 形容詞率　(g) 機能語率　(h) 接続詞率（補正後）

図 1　基礎統計量（中央値と第 1 ・ 3 四分位数）　1:白書，2:新聞, 3:小説, 4:WEB, 5:国会, 6:学会, 7:模擬

性の高低の現れとも考えられる．

第 2 の要因は文章の複雑さに関するものである．Halliday(1985) は，複雑な文章ほど動詞群の名詞化により機能語に対する内容語の比率が高くなるとし，内容語率で定義される語彙密度という尺度を提案した．佐野・丸山 (2008) は，この尺度を用いて BCCWJ の白書と書籍（文学）の比較を行い，文学よりも白書の方が語彙密度が高く，Halliday(1990) の「綿密に計画された，あるいはよりフォーマルな文章ほど語彙密度が高い」という主張と一致するとしている．機能語率の逆数が「内容語の占める割合」として語彙密度に概略相当すると考えるならば，文学，新聞，白書の順に文章としてより複雑であり，かつフォーマルということになる．

第 3 の要因は，話し言葉における和語率の高さである（野元 1959）．会話文の多い小説は話し言葉に近い側面があり，これが文学の和語率の高さにつながった可能性が考えられる．

これら三つの要因を念頭に置いて，今回追加した四つのジャンル，WEB 掲示板，国会会議録，学会講演，模擬講演の傾向を見て行こう．後者三つは話し言葉であり，また WEB 掲示板のようなインターネット上の言葉は相対的に話し言葉に近いとされることから，要因 3 を考えると，全体的に小説に近い傾向が見られることが予想される．しかし，国会会議録や学会講演については，全体的に専門性が高く，発話される文章もより複雑であり，特に前者は話し言葉としてはかなりフォーマルなものと言える．その意味では国会会議録や学会講演はより白書や新聞に近い傾向が見られる可能性もある．

さて，図 1 の結果を見てみよう．まず全体的な傾向として，追加した四つのジャンルはいずれも小説に近い傾向，つまり，漢語率，名詞率が低く，和語率，機能語率が高いという傾向を示している．しかし四つを比較してみると，予想した通り，国会会議録や学会講演は相対的に新聞や白書により近い傾向を示している．

またこの傾向が漢語率に特に強く見られることにも着目したい．確かに名詞率についても小説と比較すると国会会議録や学会講演ではより新聞や白書に近い値を示しているが，漢語率についてはそれ以上に新聞，白書に近い値を示している．これは，単に複合名詞が多いために名詞率や漢語率が高くなっているだけでなく，それに加えてこれらのジャンルでは和語よりも漢語の方が用いられやすい傾向にあることを意味する．

この結果は，本研究のような短めの言語単位ではなく，長めの単位を用いた語種比率の調査結果とも整合的である．例えば，国語研究所 (1955) では日常談話よりもニュース解説やニュースの方が，樺島 (1963) では小説よりも新聞（社会記事・新聞社説）の方が漢語率が高いとしている．この語種率の差の要因を一つに限定することはできないが，改まり度や専門性の程度の違いなどが関係しているものと考えられる．

一方 WEB 掲示板については，和語率，漢語率，名詞率に関して，小説よりは白書や新聞，国会会議録や学会講演に近い値を示しており，質問という行為においてある程度改まった言葉遣いをすることや，回答・解説における専門性の高さなどが影響したものと考えられる．

■接続詞率： 接続詞率に関してまず目につくのが，白書，学会講演，模擬講演の接続詞率の高さである．白書のような論説文や同じく論説調の学会講演で接続詞率が高いのはある程度予想できることであるが，主に個人的な内容に関するスピーチである模擬講演においてここまで接続詞率が高いというのはあまり説明がつかない．しかし実際の表現を見てみると，学会講演，模擬講演ともに，「で」が接続詞の半数以上を占めていることが分かる．国語研究所 (1955) では，話し言葉において，遊び言葉的・場つなぎ言葉的に用いられる接続詞が多いこと，実際に「それで」「だから」「で」などの接続詞が多く用いられることを指摘している．このことが，話し言葉の接続詞率を高める要因になっていると考えられる．

小椋 (2005) は，CSJ の学会講演と模擬講演の長単位の接続詞を比較し，「でも」「だから」「じゃ」のような助詞，助動詞のみで構成される接続詞が，特に模擬講演のような改まり度の低い話し言葉においてよく用いられることを指摘している．実際，今回分析対象とした七つのジャンルの接続表現を見てみると，助詞，助動詞で構成される接続詞の割合は，白書 0.1%，新聞 1.1%，小説 16.8%，WEB 掲示板 8.1%，国会議事録 2.8% であるのに対し，学会講演 64.6%，模擬講演 78.7% と，圧倒的に話し言葉に多いことが分かる．そこでこの種の接続詞を除いた上で頻度を算出し直したところ（図 1(h)），論説調の白書や国会議事録，学会講演の接続詞率が他より多いという，ある程度納得できる結果となった[*6]．

■副詞率・形容詞率： 小椋 (2005) は，CSJ を対象とした長単位の品詞率の比較調査の中で，学会講演では名詞率が高いのに対して模擬講演では副詞率や形容詞率が高いことを示し，その要因として，模擬講演における主観的表現の多さと学会講演における客観的情報伝達の多さを挙げている．また国語研究所 (1955) では，日常談話，ニュース解説，ニュースの副詞率が 6.1%, 2.5%, 1.3%，形容詞率が 2.7%, 0.9%，0.4% と，主観的表現の多い日常談話の副詞率，形容詞率が圧倒的に高いこと，また同じニュースでも，ある程度解説者の意見なども含むニュース解説の方がニュースよりも副詞率，形容詞率が高いことを示しており，小椋 (2005) の見解と一致する．

さて，図 1 を見てみると，白書，新聞，学会講演よりも小説や模擬講演の方が，副詞率，形容詞率が高い傾向が見られ，上記指摘と整合的である．特に形容詞率においてその傾向が強く見られる．副詞において，国会会議録・学会講演の副詞率が小説に比べてさほど低くないのは，接続詞の場合と同様に副詞においても，「そう」「こう」「もう」などの表現が場つなぎ言葉的に用いられているためである．なお WEB 掲示板の形容詞率が小説や模擬講演よりも高いが，これは質問の中に意見や評価を問うものが少なからずあり，その影響と考えられる．

[*6] 模擬講演の接続詞率も依然として高い．これは，模擬講演のうち，自身の過去の経験などを語るスピーチにおいて「そして」などの接続詞が多用されることがあるためである．

表 1　最適モデルの判別関数の係数と説明率 (要変更)

判別関数	第 1	第 2	第 3	第 4	第 5	第 6
漢語率	4.93	5.06	-13.86	2.07	2.75	8.30
機能語率	-21.87	7.54	-19.46	22.69	-9.20	-3.88
名詞率	2.33	1.66	7.59	18.97	7.17	-29.10
接続詞率	-2.27	183.99	115.97	56.36	-21.96	75.55
副詞率	-4.11	52.00	-35.10	-55.89	88.01	-61.96
形容詞率	2.90	-54.23	22.13	77.91	96.95	64.58
説明率	70.5%	17.9%	5.6%	4.3%	1.7%	0.0%

表 2　判別結果 (要変更)

	予測値						
観測値	白書	新聞	小説	WEB	国会	学会	模擬
白書	133	16	0	0	0	1	0
新聞	15	123	9	1	1	1	0
小説	0	2	137	7	2	0	2
WEB	0	6	27	93	15	4	5
国会	0	0	0	0	150	0	0
学会	3	9	0	4	11	97	26
模擬	0	1	12	12	8	11	106

4　ジャンルの判別

このように各品詞率，語種率に関して，ジャンル毎に異なる使用傾向が見られることがわかった．そこで本節では，線形判別分析を用いて，これらの説明変数から当該文章のジャンルを推定するモデルを構築し，各変数がジャンルの判別にどのように寄与するか，また各ジャンルがこれらの変数から見てどの程度類似しているかを検討する．説明変数は前節で着目した指標に限定した．

分析には R の MASS パッケージに含まれる lda 関数を用いた．ステップワイズ変数選択（変数増減法）で最適なモデルを求めた．その結果，七つの変数のうち和語率を除く六つが選択された．最適なモデルとして選択された判別関数の係数と説明率を表 1 に示す．また図 2 に，横軸を第 1 判別関数，縦軸を第 2 判別関数とした散布図を示す．最適モデルを用いた leave–1–out 交差検証を行ったところ，正解率は 79.9% であった．判別結果を表 2 に示す．

第 1 判別関数の係数を見ると（表 1），正の値として漢語率が，負の値として機能語率が大きな値を示している．このことから，前節の議論に従えば，第 1 軸は概ね専門性や改まり度の程度，あるいは書き言葉的／話し言葉的なものを分ける軸と考えられる．図**??**の横軸方向を見てみると，正から順に，白書，新聞，小説・WEB 掲示板・話し言葉が並んでおり，また細かく見ると，国会会議録や学会講演がより新聞に近い位置にプロットされてり，確かにこの種の要因が関係した軸であることがうかがえる．

また新聞に関して第 1 軸の方向に分散が大きいのは，一般記事からコラムまで幅広くサンプルをとっているためと考えられる．WEB 掲示板の分散もかなり大きいが，実際にデータを見ても，特に質問に対する回答の文体がかなりくだけた（あるいはふざけた）ものから，丁

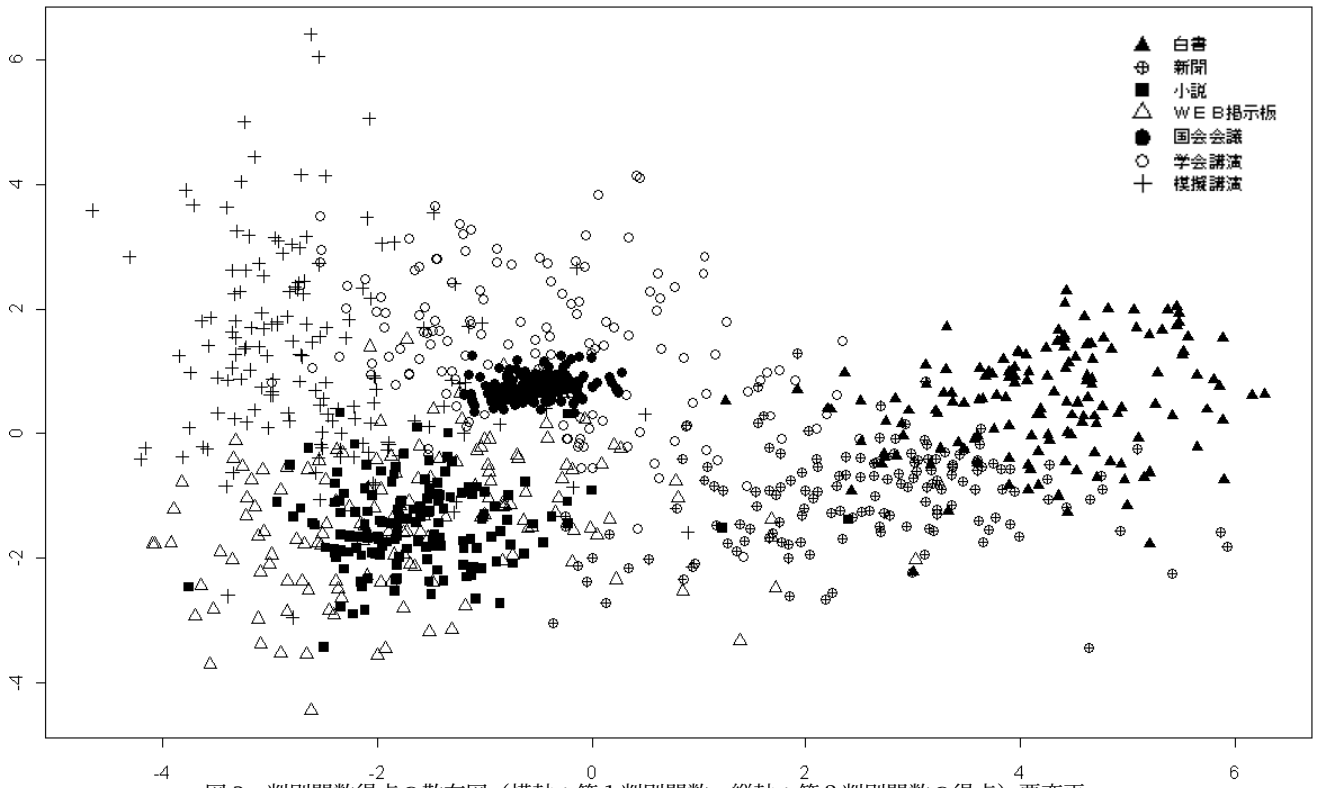


図 2　判別関数得点の散布図（横軸：第 1 判別関数，縦軸：第 2 判別関数の得点）要変更

寧かつ詳細に解説をしているものまで多様であり，その現れと考えられる．学会講演も若干横軸方向に分散が大きいが，これも学会によって改まり度に差があること，また読み上げ原稿のある書き言葉に近い話し方のものから自発性のかなり高いものまで幅が広いことの影響と考えられる．

逆に，国会会議録の分散がかなり小さく，学会講演のある一箇所に固まっていることも印象的である．図 1 から分かるように，国会会議録ではどの指標を見ても分散が極めて小さい．これは，国会での会議のスタイルや内容がある程度固定していること，更に会議録が CSJ のような話し言葉の忠実な書き起こしではなく，ある程度手が加えられていることに因るものと考えられる．

次に第 2 判別関数の係数を見てみよう．ここでは圧倒的に接続詞が正の方向に効いていることが分かる．実際，図 1 の接続詞率の分布がそのまま図 2 の第 2 軸方向に展開された形になっており，正の方向に学会講演や模擬講演，国会議事録，白書が，負の方向に新聞，WEB 掲示板，小説が配置されている．しかし前節でも指摘したように，場つなぎ言葉的な接続詞なども含まれており，話し言葉と書き言葉の比較を単純に接続詞率あるいは副詞率で見ることは妥当ではない．前節では助詞・助動詞のみで構成される接続詞を話し言葉で用いられやすいものと仮定して分布を見た．実際，助詞・助動詞のみで構成される接続詞についてはかなり良い指標と言えそうであるが，それで全てが尽くされたわけではないことは，補正後の模擬講演の接続詞率の分布からも明らかである．また副詞についても同様に検討の余地がある．これらについては今後の課題としたい．

謝辞：線形判別分析に関して千葉大学の伝康晴氏の協力を得ました．ここに記して感謝します．

参考文献

小椋 (2005) 「『日本語話し言葉コーパス』の資料性—形態論情報を用いた分析から—」『国語語彙史の研究』24, 259–275, 和泉書院.

小椋 (2007) 「『日本語話し言葉コーパス』の語種構造」『話し言葉コーパスに基づく言語変異現象の定量的分析』科学研究費補助金研究成果報告書.

小椋ほか (2008a) 『『現代日本語書き言葉均衡コーパス』形態論情報規定集』国立国語研究所内部報告書.

小椋ほか (2008b) 「形態素解析用辞書への語種情報の実装と政府刊行白書の語種比率の分析」『言語処理学会第 14 回年次大会発表論文集』935–938.

樺島 (1963) 「漢語をめぐって」『計量国語学』27, 14–19.

小磯ほか (2008a) 「『現代日本語書き言葉均衡コーパス』にもとづくジャンル間の文体差に関わる要因の分析」,『社会言語科学会第 22 回研究大会発表論文集』192–195.

小磯ほか (2008b) 「短単位情報に基づくジャンル間の文体に関する分析」『特定領域研究「日本語コーパス」平成 20 年度全体会議予稿集特定領域研究』99–106.

国語研究所 (1955) 『談話語の実態』国立国語研究所報告 8, 秀英出版.

佐野・丸山 (2008) 「システミック文法に基づく書きことばの複雑さ測定– 日本語大規模コーパスを用いた語彙密度計測–」『言語処理学会第 14 回年次大会予稿集』, 1097–1100.

野元 (1965) 「話しことばの中での漢語使用」『ことばの研究』国立国語研究所論集 1 .

Halliday(1985) *Spoken and Written Language*. Victoria: Deakin University.

Halliday(1990) Some grammatical probalems in scientific English, *Annual Review of Applied Linguistics*, 6, pp.13–37.