長単位に基づく『現代日本語書き言葉均衡コーパス』の品詞比率に関する分析

冨士池優美　　小西光　　小椋秀樹　　小木曽智信　　小磯花絵

人間文化研究機構　国立国語研究所

1.　はじめに

『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）には「コア」[1]と呼ばれるデータセットがあり，自動解析結果を人手修正した精度の高い「短単位」「長単位」情報が提供される。各サンプルには文章の内容を表すカテゴリ情報[2]が付与されている。

本発表では，長単位情報を利用し，コアデータのうち，中央官庁刊行の白書，書籍，新聞，雑誌，Yahoo!知恵袋（以下，知恵袋）を対象に品詞比率を調査し，サンプルの掲載媒体とカテゴリ情報の二つの観点から，文体との関係について検討する。

2.　長単位の概要

長単位は，構文的な機能に着目し，文章の言語的特徴の解明を目的とした言語単位である。

長単位では「国立国語研究所」「予備的分析」「表示する」のような複合語を1単位として認める。「だ」「を」のような付属語は単独で長単位とするのが原則であるが，「ので」「ている」のような複合辞も付属語として1長単位としている[3]。

長単位の品詞情報は，文脈に即して品詞を付与する。短単位に付与されている名詞-普通名詞-形状詞可能，名詞-普通名詞-副詞可能などは，その用法に基づき，名詞・形状詞・副詞に判別している。「結果」を例とすると，「これらの結果に基づき」の場合は名詞を，「結果、様々な社会問題が発生し」の場合は副詞を付与する。

3.　品詞比率

ここでは白書・書籍・新聞・雑誌・知恵袋の長単位コアデータを調査対象とする。表1に長単位コアデータの延べ語数を示す。資料規模の参考として，短単位延べ語数をあわせて示した。

表1　長単位コアデータ延べ語数

	白書	書籍	新聞	雑誌	知恵袋
長単位	159021	195333	273878	200294	95110
短単位	228272	229723	360825	245540	110691

[1] 「コアデータ」の設計については小椋ほか（2009）を参照。

[2] カテゴリ情報は，BCCWJにおいて「ジャンル情報」として付与されている。詳細については丸山（2009）を参照。

[3] 認定基準の詳細については小椋ほか(2011)を参照。

3.1　先行研究

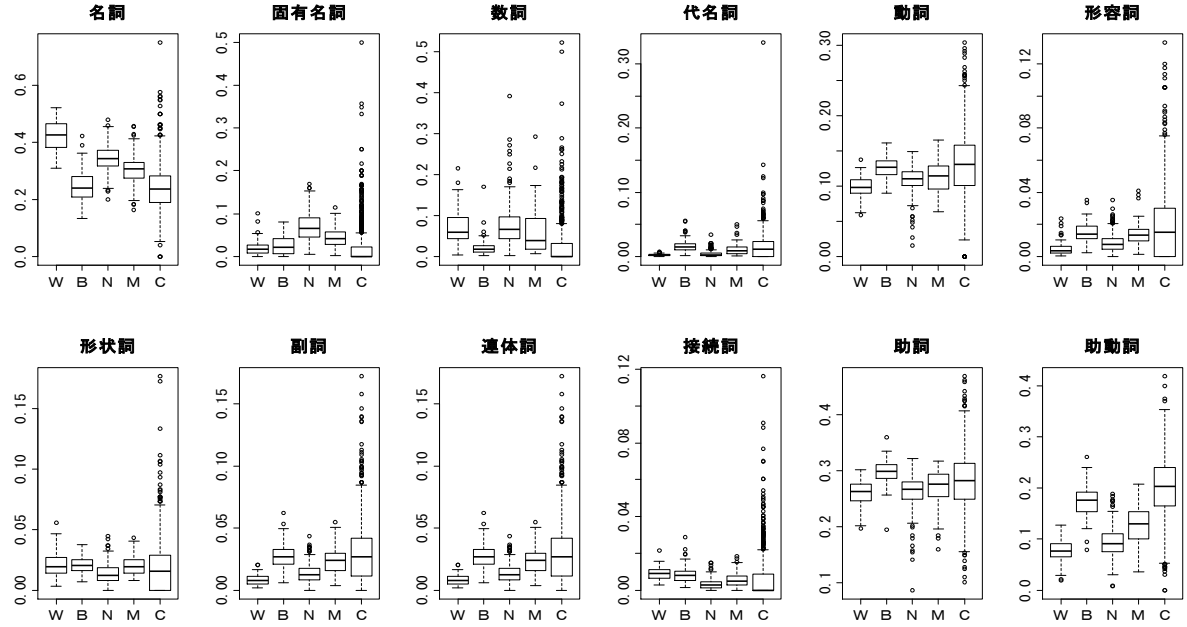
長単位に基づく品詞構成に関する研究は，これまでに小磯ほか（2009），冨士池ほか（2010）がある。小磯ほか（2009）は白書・新聞記事・社説・小説と講演を対象とした調査，冨士池ほか(2010)は白書・書籍・新聞の長単位コアデータを対象とした調査である。これらで分析対象となった長単位データは，名詞-普通名詞-〇〇可能が未判別であり，2章で挙げた「結果」の例にはどちらも名詞が付与されている。今回は，名詞・形状詞・副詞の判別を行い精密化したデータを利用し，対象の媒体を増やして分析を行う。

3.2　媒体別品詞比率

品詞比率（空白・記号・補助記号・URL類を除く，延べ語数）の基礎統計量を媒体別に示したものが図1である。「名詞」は固有名詞・数詞を除いたものである。

名詞率は媒体差が大きく現れており，知恵袋，書籍，雑誌，新聞，白書の順に高くなっている。動詞率は，知恵袋・書籍と比較して，新聞・白書で比率が低くなっており，雑誌はその中間と，名詞率と負の相関にある。形容詞・副詞・連体詞といった相の類の比率も動詞率と同様に，名詞率と負の相関にある。相の類の中で形状詞率のみ傾向が異なり，媒体差が小さく，新聞で比率がやや低くなっている。固有名詞率は新聞で高く，数詞率は書籍・知恵袋で低く，代名詞率は書籍で比率が高くなっており，媒体ごとの内容の特徴が反映されたものと考えられる。助詞率は媒体差が小さく，書籍でやや比率が高くなっている。助動詞率は白書，新聞，雑誌，書籍，知恵袋の順で高く，動詞や相の類と同様に，名詞率と負の相関にある。また，知恵袋は他媒体と比較してサンプルの分散が大きく，品詞の別なく，極端に比率の高いサンプルがあることがわかる。

冨士池ほか（2010）と今回の結果を比較してみよう。書籍の形状詞率，新聞の副詞率がより高くなり，白書は変化が小さかった。これらは，用法に基づき名詞・形状詞・副詞の判別をした結果，媒体の特徴がより明確になったものと考えられる。



W:白書, B:書籍, N:新聞, M:雑誌, C:知恵袋

図 1　品詞比率

相の類の比率は名詞率と負の相関関係を持つが, 形状詞率のみ傾向が異なることを先に述べた。白書の名詞率は他媒体より際立って高く, 負の相関関係がある相の類の一つである形状詞率は他媒体よりも低くなることが予想されるが, 実際には白書の形状詞率は他の媒体と同程度であり, 予想より高い。冨士池ほか（2010）では形状詞的接尾辞「的」の頻度が白書で高いことを示したが, 名詞の形状詞化により形状詞率を高めている可能性がある。

3.3　媒体と文体の関連

ここで, 体・用・相の三つの類の相関を見たい。樺島・寿岳（1965）は「100×相の類の比率／用の類の比率」で求められる MVR という指標を提案し，MVR と名詞の比率との組み合わせから，名詞の比率が大きく MVR が小さければ要約的な文章，名詞の比率が小さく MVR が大きければありさま描写的な文章，名詞の比率が小さく MVR も小さければ動き描写的な文章と考えられるとしている。

図 2-1 は，体の類（%）に対する MVR の分布である。体の類の比率（%）を x 軸，MVR を y 軸にとっている。

図から，他の媒体と比較して，知恵袋は分散が大きいことがわかる。体の類が低く MVR が小さい，動き描写的なものと，体の類が低く MVR が大きいありさま描写的なものが多いが, 体の類が

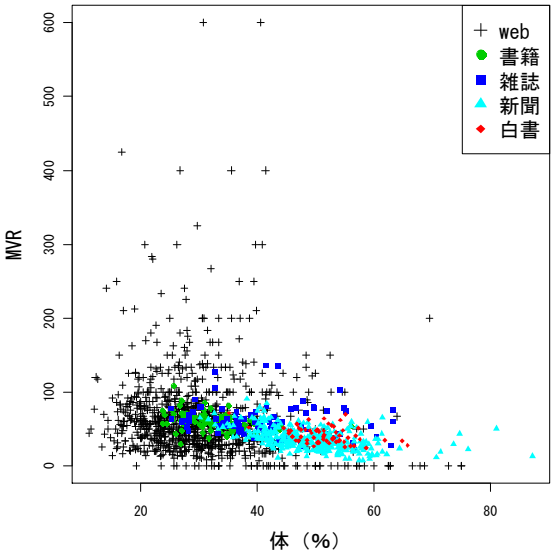


図 2-1　体の類（%）に対する MVR の分布

極端に高い, 要約的なものも少ないながら観察され，文体が一様でないことがわかる。

サンプルの例を以下に示す。

体：大，MVR：小

ＪＮＢ→郵便局口座への振込み手数料はいくらでしたっけ？？？ ジャパンネットバンクの「郵貯Ｗｅｂ送金」のことですよね？振込手数料は２９４円です。

（知恵袋，体：62.9%，MVR:0）

◆サッカーＪ２第３６節　コンサドーレ札幌―湘南ベルマーレ　２７日午後２時、札幌ドーム（豊平区羊ケ丘１）。前売りＳＳ席４２００円、Ｓ指定席３７００円、ＳＡゾーン席３千円（小中学生千円）、ＳＢゾーン２５００円（同８００円）、Ｂ自由席２千円（同６００円）。当日券各２００円増し。北海道フットボールクラブ☎０１１・７５０・２９３６

（北海道新聞，体：87.2%，MVR：12.9）

体：小，MVR:大

クレヨンしんちゃんをみたのですが
しんちゃんは、ぴちぴちおねいさんが大好きですが実際の５歳児もしんちゃんのようにおねいさんが好きなのでしょうか
嫌いではないでしょう。むしろ大好きでしょう。ただ、しんちゃんみたいに、積極的かどうか疑問です男の子だから、しょうがないと言ってしまえばそれまでですが・・・

（知恵袋，体：30.7%，MVR:600.0）

体：小，MVR：小

ニンニクが臭いというのは消化して食道の中から出てくる臭いですよね？
そのものをこんがりあぶると香ばしいのですが臭いという人がいます。にんにくの臭いは、消化して、血液の流れに乗り、肺にたどり着くのです。そして、呼吸とともに臭いがでてくるのです。

（知恵袋，体：20.9%，MVR：50.0）

次に白書・新聞・雑誌・書籍が集中する部分を見てみよう。図 2-1 のうち，知恵袋を除いたものが図 2-2 である。概ね，書籍，雑誌，新聞，白書の順に体の類の比率が大きくなり，これに従い MVR が小さくなるのが見てとれる。要約的な文章と考えられるものに新聞と白書があり，新聞には極端に体の類の比率が高いものがある。これに対して，書籍・雑誌はありさま描写的な方向に分布している。特に雑誌はありさま描写的な傾向が強いが，要約的なものもあり，分散が大きい。

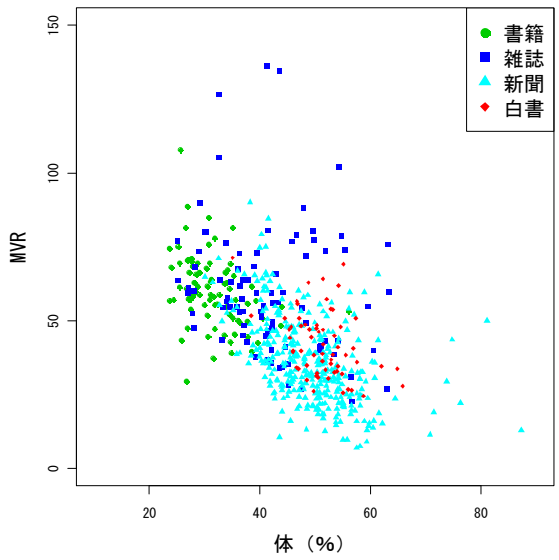


図 2-2　体の類（%）に対する MVR の分布（知恵袋を除く）

サンプルの例を以下に示す。

MVR：大

カジュアルもキレイめもお手のもの！　かわいらしさが残るカジュアルなスタイリングは、マネしたいポイントがいっぱい。
ｓｃｅｎｅ　デート
デートの日は絶対ミニスカ！　フワフワファーで女のコらしく

（雑誌 CanCam，体：41.5%，MVR：136.2）

3.4　カテゴリ・媒体と文体の関連

BCCWJ では，各サンプルに文章の内容を表すカテゴリ情報が付与されている。具体的には，書籍の日本十進分類表（NDC），雑誌における『雑誌新聞総かたろぐ』の「分野」，新聞の配達エリア（全国紙・ブロック紙・地方紙），知恵袋の質問が投稿されたカテゴリ名などである。白書については，タイトルの内容に応じて，国立国語研究所で独自に分類したものが付与されている。

品詞構成と文体の関連の研究においては，観点として形式（新聞の記事・社説，小説，短歌・俳句等）を設定することが多く，国立国語研究所の語彙調査では媒体（放送，雑誌等）を観点とするが，カテゴリ（内容）も文体に影響している可能性がある。3.3 節まで見てきた媒体差についても，各媒体に含まれるカテゴリの偏りが擬似的に媒体差として現れた可能性もある。そこで，カテゴリを限定した上で同じ分析を行い，全体の場合と比較し，3.3 節と同様の媒体差が現れるかを確認する。また，カテゴリの文体に与える影響についても検討する。

各媒体に共通するカテゴリとして，書籍の日本十進分類法 3 番台（**社会科学**）を中心に，雑誌・白書・知恵袋についてはその下位分類（社会科学，政治，法律，経済，財政，統計，社会，教育，風俗習慣．民俗学．民族学，国防．軍事）[4]と共通・類似した名称を持つもの[5]を選定し，以下の 4 媒体 9 サブカテゴリを対象とした。新聞のカテゴリ情報「配達エリア」からは内容が判別できないため，新聞は除外した。

- **書籍**　社会科学　…18 サンプル
- **雑誌**　政治・経済・商業　…7 サンプル
- **白書**　…30 サンプル
  - 安全
  - 外交
  - 教育
  - 経済
- **知恵袋**　…129 サンプル

[4] NDC 新訂 9 版分類表（2 次区分表）による。
[5] 雑誌には「教育・学芸」のカテゴリがあるが，これは文芸雑誌の小説・批評に付与されているため，除外した。

ビジネス、経済とお金
ニュース、政治、国際情勢
子育てと学校

これらの社会科学系サンプルについて，体の類に対する用・相の類の割合の関連を見るために，体の類の比率(%)と MVR の分布を図 3 に示す。

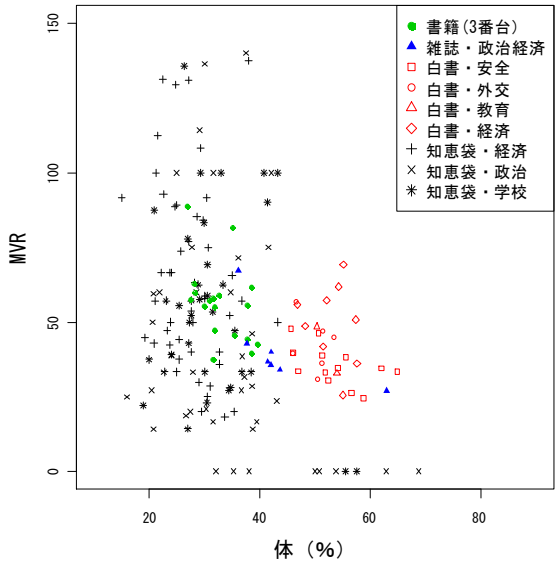


図 3　体の類（%）に対する MVR の分布（社会科学系）

図 3 を見ると，全体の傾向を示した図 2-1 と同種の傾向，例えば白書は体の類の比率が高く，知恵袋は MVR が小さいものもあれば大きいものもあり分散が大きい[6]ことが見てとれる。このことから，図 2-1 で見た媒体差は，各媒体に含まれるカテゴリの偏りが擬似的に現れたものではないことがわかる。

細かく見ると，白書と知恵袋については全体の場合とほぼ同様の分布となっているのに対し，雑誌については全体とは若干異なる傾向が見られた。社会科学系に限定した結果，体の類の比率はほぼ差がなかったが，MVR は全体で 22.4 から 136.2 まで分布していたものが社会科学系では 27.0 から 67.2 と低い位置での分布となっており，形容詞・副詞の類が抑制されている傾向が見られた。また，書籍についても同様に，MVR が低く，相の類が抑制される傾向が見られた。

このように社会科学というカテゴリでは，全体と比べた場合に MVR が相対的に低くなる，つまり形容詞・副詞の類が抑制される傾向が，雑誌と書籍に共通して見られることから，媒体とは別にカテゴリが品詞構成に影響を与えている可能性のあることが示唆される。白書や知恵袋などではこの影響は観察されなかったが，白書については行政報告書という媒体自体の制約が強いことに起因している可能性がある。知恵袋は，個人的な経済状況の相談など，社会科学系ではないサンプルが含まれていることに起因していると考えられる。

[6] 図 3 は y 軸を 0 から 150 までに取っているが，150 から 600 の間に知恵袋 7 サンプルが分布している。

4.　まとめ

媒体別の品詞比率と，サンプルを社会科学系に絞った場合の品詞比率から，①名詞と動詞・形容詞・副詞・連体詞・助動詞の比率は負の相関関係にある，②白書・新聞は書籍・雑誌と比較して要約的，雑誌はありさま描写的な傾向が強く，知恵袋は品詞比率の分散が非常に大きい，③カテゴリを限定しても媒体による品詞構成差が見られる一方で，媒体によってはカテゴリを限定することで全体と異なる傾向が共通して見出されたことから，カテゴリが品詞構成に影響を与えている可能性があるということがわかった。

雑誌という中間的な媒体が増えたことで，新聞の品詞構成の特徴がより明確になり，知恵袋については文体が一様ではないことが明らかになった。コアデータには web データとして Yahoo!ブログも収録される。これも合わせて分析することによって，さらに web における媒体差が明確になることが期待される。

参考文献

小椋秀樹ほか（2009）　「『現代日本語書き言葉均衡コーパス』における形態論情報付与作業の進捗状況」『特定領域「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集』pp.57-64
小椋秀樹ほか(2011)　国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版』
樺島忠夫・寿岳章子（1965）　『文体の科学』(綜芸社)
小磯花絵ほか（2009）　「長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「日本語コーパス」平成 21 年度公開ワークショップ（研究成果報告会）予稿集』pp.183-190
冨士池優美ほか（2010）　「『現代日本語書き言葉均衡コーパス』長単位情報に基づく予備的分析」　『特定領域「日本語コーパス」平成 22 年度全体会議予稿集』pp.101-108
丸山岳彦（2009）　「『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）サンプリング方法について」(『現代日本語書き言葉均衡コーパス』モニター公開データ（2009 年度版）DVD 所収)

付記　本研究は，文部科学省科研費特定領域研究「日本語コーパス」による補助を得た。