一般的な日本語テキストにおける助詞比率の規則性

一般的な日本語テキストにおける助詞比率の規則性

森　秀明（東北大学文学研究科）†

Regularity of the Particle Ratio in General Japanese Texts

Mori Hideaki (Graduate School of Arts and Letters, Tohoku University)

要旨

日本語のテキストでは名詞比率に連動して動詞や形容詞などの比率が規則的に変化することが知られている。しかし名詞比率と付属語の関係は明らかにされていないため,『現代日本語書き言葉均衡コーパス』(以下 BCCWJ) 固定長・長単位データと『日本語歴史コーパス』(以下 CHJ) の長単位データを使用し，名詞比率と助詞比率の相関を中心に観察した。BCCWJ の中には，例えば商品名と値段が列挙されるなど，ほとんど助詞が使用されないサンプルが存在するため,「名詞比率 45%未満・その他比率 30%未満」のサンプルを仮に「一般的な日本語テキスト」と定義して調査した。この結果，連体助詞には名詞比率と正の相関が，接続助詞には負の相関があるなど，様々な相関が認められた。また注目すべきことに助詞の中分類ではこのように名詞比率との相関がありながら，それらを合計した大分類では，多くのテキストの助詞比率は 34%前後とほぼ一定で, その比率は古典語でも同じであった。

１．研究の目的と先行研究

日本語のテキストで使用されている品詞の構成比率には一定の規則性が存在し，名詞比率に連動して動詞や形容詞類の割合が規則的に変化することが知られている。樺島（1955）は現代語の延べ語数を使用した単位語水準の品詞構成比率（図 1）を，大野（1956）は古典文学の異なり語数を使用した見出し語水準の品詞構成比率（図 2）を分析し，これを明らかにした。図 1 のマーカーは名詞比率の低いものから日常会話，小説会話，哲学書，小説地の文，自然科学書，和歌，俳句，新聞記事の順となっており，名詞の増加は話し言葉から書き言葉へ，文の凝縮度の低いものから高いものへと向かっている（図中の線は樺島，1955:55 の数式に基づく)。図 2 のマーカーは同じく源氏物語，竹取物語，讃岐典日記，紫式部日記，土佐日記，枕草子，方丈記，徒然草，万葉集で，物語，日記，随筆など同じジャンルの作品が似た品詞比率になっており，図 1 と同様の傾向が観察される。

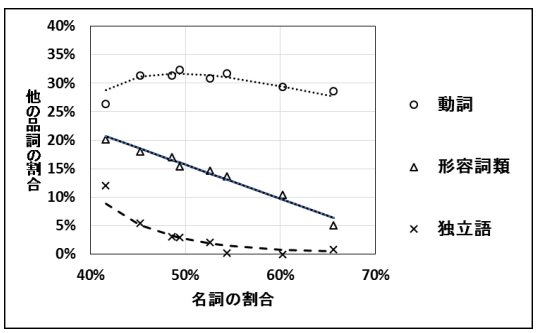

図 1：樺島（1955）第一表に基づく散布図

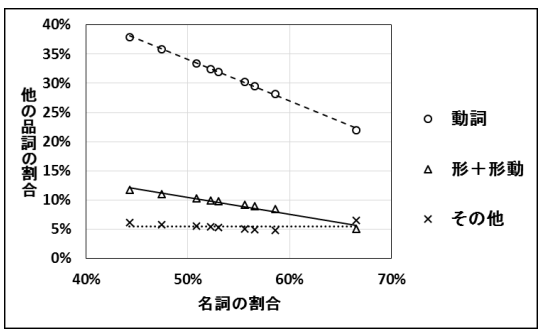

図 2：大野（1956）第七表に基づく散布図

† hideaki@moriharuo.com

図 1，2 に見られる規則性を定式化した数式は「樺島の法則」や「大野・水谷の法則」と呼ばれ，計量的な言語研究における重要な発見と位置づけられてきた。ただしこれらの研究によって名詞と自立語の関係については明らかになったが，名詞と付属語の関係は不明なままである。付属語も含めた日本語の品詞比率の研究には冨士池ほか（2011）や山崎（2014）などがあるものの，名詞との相関は調査されていない。そこで本研究では BCCWJ の固定長・長単位データと CHJ の長単位データを使用し，付属語の中でも助詞に焦点を当て，名詞比率との相関を中心にその規則性を観察する。

２．分析データ

２.１　使用するコーパスとデータの種類

分析には国立国語研究所が公開している BCCWJ の固定長･長単位データと CHJ の平安，鎌倉，室町編の長単位データを使用する。ただし BCCWJ 固定長データの図書館書籍サブコーパス（以下図書 SC と略す）と，書籍サブコーパス（以下書籍 SC と略す）の分析結果はよく似た結果となったため，本研究では図書 SC の結果のみ提示する。BCCWJ と CHJ では形態素解析用辞書 UniDic と長単位解析器 Comainu によって品詞情報が付与されている。UniDic の品詞体系は基本的に学校文法の体系に近いが，形容動詞はその語幹を「形状詞」として認定され，活用語尾は助動詞に分類されている。また長単位では複合名詞を 1 語に認定するほか，「における」「という」「である」などの複合助詞，複合助動詞を一語として認定している。本研究では格助詞ノを連体助詞として格助詞から分離して分類する以外，品詞の認定は UniDic の品詞体系に従った。また本研究では品詞を類別して分析する際，基本的に山崎（2014）の類別基準を参考にしたが，品詞比率が大きい名詞，動詞，助詞，助動詞以外は一括して「その他」として扱った[1]。また格助詞や係助詞と言った助詞の下位分類を中分類，それらを合計した助詞全体を大分類と呼ぶ。

２.２　データの絞り込み

図 1 は，BCCWJ 図書館書籍（以下図書と略す）の 10,551 サンプルについて品詞比率を求め，横軸を名詞比率，縦軸を助詞比率にして描いた散布図である。

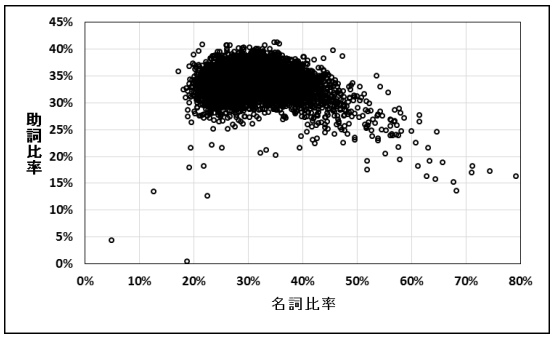

図 1 名詞比率と助詞比率の散布図：
図書 SC，N=10,551

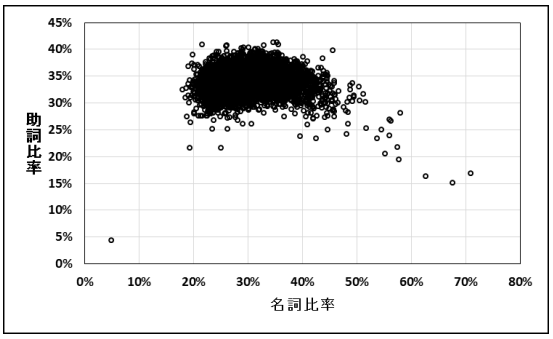

図 2 名詞比率と助詞比率の散布図：
図書 SC 章節構造文書，N=8,792

[1] 名詞：名詞・代名詞・接尾辞－名詞的，動詞：動詞，接尾辞－動詞的，助詞：助詞，助動詞：助動詞，その他：長単位語数表（BCCWJ_WC_LUW_v10.xlsx）の語数（記号等除外・固定長）から上記の品詞数を除いたもの。山崎（2014）では名詞に「記号」を含めるが，本研究では「その他」の品詞数の算出に長単位語数表（記号等除外・固定長）を使用したため，名詞に「記号」は含めなかった。

図 1 では名詞比率 40%までは楕円形で，そこから下に向かう尾がついているような形をしている。図 2 は国立国語研究所（2015）の文体情報[2]を使用し，柏野（2013）で「文体判断が単純にいかないもの」と判断された 1,758 サンプルを除いた上で図 1 と同様に描いた散布図である。「文体判断が単純にいかないもの」は図解，コマ割などが多用される「視覚表現多用系」，用語解説，見本・カタログ形式などの「データベースやリスト系」，対談，インタビューなどの「対話系」など 11 の観点から分類されているサンプルで，「テキスト構造・紙面形式に特徴をもつもの」である。次の（1）は「視覚表現多用系」，（2）は「データベースやリスト系」の文書の一部である。

（1）アリのなかまクロオオアリアリ科■働きアリ７～十三ｍｍ■４～十月　全国■里山■成虫・幼虫●日本では最大のアリ働きアリ女王アリ←ムネアカオオアリアリ科■働きアリ８～十二ｍｍ■５～十月■北・本・四．九■里山■成虫・幼虫●クロオオアリに似るが胸が赤い（BCCWJ サンプル ID：LBqn_00015，実著者不明，『昆虫』，名詞比率 50.9%，助詞比率 27.5%）

（2）今後，世界遺産条約の締約が期待される中東の国々アラブ首長国連邦Ｕｎｉｔｅｄ　Ａｒａｂ　Ｅｍｉｒａｔｅｓ面積　八万三千六百ｋｍ２人口　二百五十八万人主要言語　アラビア語首都　アブダビ通貨　ディルハム民族　アラブ人宗教　イスラム教（BCCWJ サンプル ID：LBo5_00063，実著者不明，『世界遺産ガイド』，名詞比率 71.4%，助詞比率 18.2%）

（1），（2）の文書では助詞の数に比べ名詞の数が著しく多い。その理由はこれらの文書に項目のリストとして名詞句の列挙が多く含まれるからである。これらの「文体判断が単純にいかないもの」を除くと，図 2 のように尾の部分がかなり少なくなる。それでもまだ図 2 では名詞比率 45%までの楕円形の塊と尾に分かれているように見える。

次に図 2 の尾の部分のサンプルを観察する。（3）は図 2 で最も名詞比率が高いサンプル（4）は名詞比率 44.5%のサンプル（5）は名詞比率が最も少ないサンプルである。

（3）また，高速十号線（新宿区付近～練馬区付近），同内環状線（墨田区付近～新宿区付近同十一号線（葛飾区付近～市川市付近），同晴海線（江東区付近～千代田区付近），同磯子線（横浜市南区付近～同市磯子区付近），同２号線（延伸），第二東京湾岸道路，都心新宿線及び首都高速道路４号線の機能強化について計画を進める。（BCCWJ サンプル ID：LBg6_0001，実著者不明，『首都圏白書』，名詞比率 70.9%，助詞比率 17.0%）

（4）宗室は有爵と無爵があり、爵位は次の十四等に別れる。　親王、世子、多羅郡王、長子、多羅貝勒、固山貝子、鎮国公、輔国公、不入八分鎮国公、不入八分輔国公、一・二・三等鎮国将軍、一・二・三等輔国将軍、一・二・三等奉国将軍、奉恩将軍。　（BCCWJ サンプル ID：LBi9_00142，高陽（著）永沢道雄・鈴木隆康（訳）『西太后』，名詞比率 44.5%，助詞比率 35.2%）

（5）２、無政府主義派（イ）共産主義ノ主張ハ基礎ヲ社会大衆ニ置キ、巧ミニ之レヲ誘致

[2] 柏野（2013）は図書 SC のサンプルに「専門度」「客観度」「硬度」などの文体指標を付与した研究の紹介論文で，その成果を公開しているのが国立国語研究所（2015）である。

シテ民衆的革命ヲ目的トスルニ反シ、無政府主義ハ権力ヲ否定シ、暴力革命ヲ高調スル点ニ於テ今次ノ如キ突発事変ニ際シテハ警戒ノ必要寧ロ前者ヨリ以上必要トスルモノアリ。（BCCWJ サンプル ID：LBs2_00005，松尾尊兌，『世界史としての関東大震災』，名詞比率 4.8%，助詞比率 4.4%，その他比率 87.1%）

（3）は柏野（2013）で「文体判断が単純にいかないもの」には認定されていないが，道路の名前が列挙されており，一般的なテキストとは見なしにくい。（4）も後半は名詞の列挙で一般的な文章になっていない。（5）は名詞がたくさん出現しているが，名詞比率は 4.8%となっている。その理由はほとんどの品詞を「カタカナ文」というカテゴリで解析されているためで，うまく形態素解析できていないと考えられる。本研究の目的は名詞比率と助詞比率の相関を観察することにあるため，これらのサンプルを含めて観察する意味は小さい。よって本研究では名詞の列挙が多く含まれるサンプルや（5）のようなカタカナ交じり文が多く含まれるサンプルがなるべく含まれないような絞り込みを行う。図書 SC 以外の固定長データには，国立国語研究所（2015）のような文体情報を付与した研究が存在せず，（3）～（5）に見られるように，国立国語研究所（2015）の文体情報を使用しても必ずしも本研究の目的にふさわしいサンプルに絞り込めるとは限らない。そこで名詞の列挙を含む文を少なくする目的で名詞比率は 45%未満に，「カタカナ文」を多く含む文書を少なくする目的でその他比率は 30%未満に絞り込み，この「名詞比率 45%未満・その他比率 30%未満」のサンプルを仮に「一般的な日本語テキスト」と定義してこれを分析に使用する。

絞り込みの結果，図書 SC では全体の 98.2%に当たる 10,364 サンプルが残った。図 3 はこのデータを使用した名詞比率と助詞比率の散布図，図 4 は名詞と助詞の度数折れ線である。助詞比率の平均は 34.1%，標準偏差 2.1%，名詞比率の平均は 29.7%，標準偏差 4.5%である。助詞は 34.1%±2.5%の範囲に 8 割のテキストが存在し，非常に狭い比率の範囲で使用されている。また名詞の比率との相関はない（決定係数 R^2＝.001）。

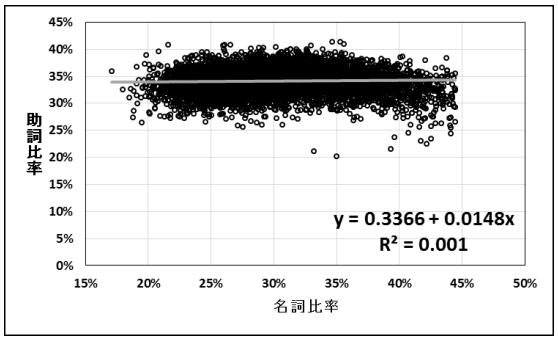

図 3 名詞比率と助詞比率の散布図：
一般的な日本語テキスト，N=10,364

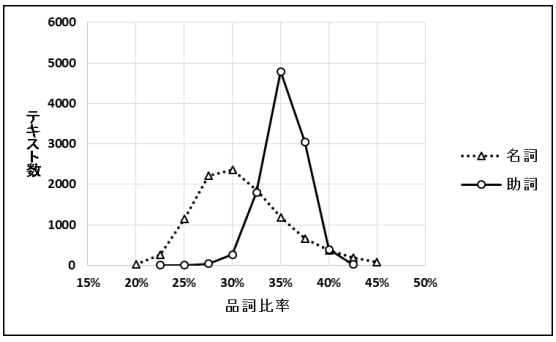

図 4 名詞と助詞の度数折れ線：
一般的な日本語テキスト，N=10,364

３．分析結果①：名詞と助詞の大分類との相関

前節では，図書 SC のサンプルの中には名詞の列挙や形態素解析の不具合によって，名詞と他の品詞との相関を観察するのが難しいサンプルが存在することを述べた。またこれらを除く目的で名詞比率 45%未満・その他比率 30%未満の文書に絞り込むと，名詞と助詞には相関がなく，助詞が 34.1%±2.5%の狭い範囲で使用されているサンプルが多いことが分かった。

本節では BCCWJ の図書 SC，新聞 SC，雑誌 SC，白書 SC の固定長・長単位データと CHJ の平安，鎌倉，室町編の長単位データから名詞比率 45%未満・その他比率 30%未満のサンプルを絞り込んだ「一般的な日本語テキスト」を使用して，助詞の大分類による規則性を中心に観察する。図 5〜9 は一つ一つのテキストに対し，名詞率，動詞率，その他率，助動詞率，助詞率を求め，名詞比率の昇順にソートして棒グラフを描いた図である。面のように見えるが棒グラフが大量に連なっている。これを見ると，助詞以外の品詞は名詞の比率が高くなるのに連動して比率が低くなるが，助詞はほぼ一定で変化しないことが分かる。

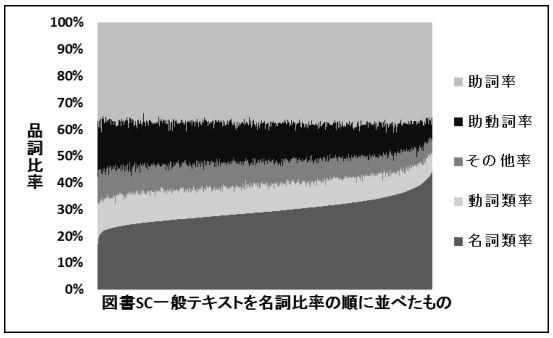

図 5 図書 SC の品詞比率，N=10,364

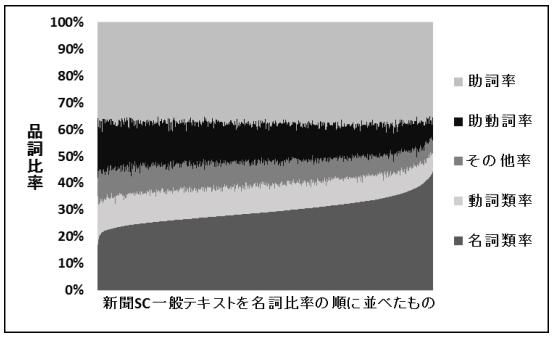

図 6 新聞 SC の品詞比率，N=1,473

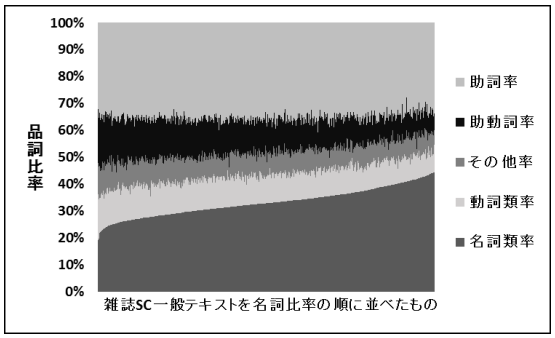

図 7 雑誌 SC の品詞比率，N=1,690

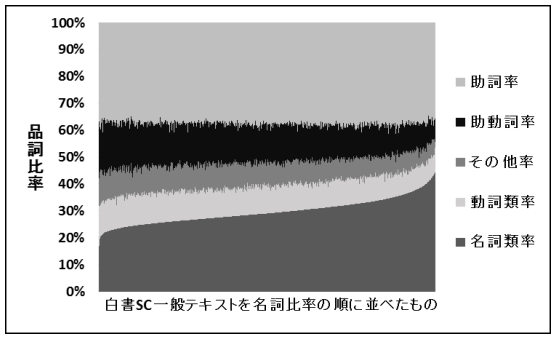

図 8 白書 SC の品詞比率，N=1,147

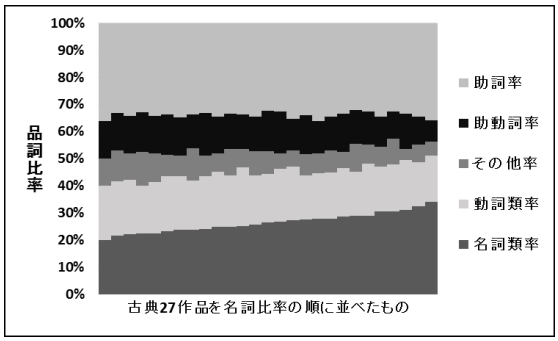

図 9 CHJ27 作品の品詞比率，N=27

表1 助詞比率の基本統計量

	図書SC	新聞SC	雑誌SC	白書SC	CHJ
平均	34.1%	34.8%	33.3%	34.2%	33.9%
標準偏差	2.1%	2.0%	2.6%	2.1%	1.1%
最小	20.3%	27.4%	22.0%	27.6%	32.1%
最大	41.4%	40.3%	40.7%	40.7%	36.1%
尖度	1.209	−0.09	0.965	−0.03	−0.5
歪度	−0.4	−0.29	−0.71	0.07	0.382
標本数	10364	1333	1690	1147	27
全サンプル	10551	1473	1996	1500	27
残存率	98.2%	90.5%	84.7%	76.5%	100.0%

表 2 は名詞比率を説明変数，4 種類の品詞比率を目的変数とした回帰分析を行って求めた回帰直線の傾きと切片，および決定係数 R^2 の値で，図 10〜14 はこれを図示したものである。雑誌 SC や白書 SC では，名詞比率と助詞比率に弱い負の相関が観察されるが，図書 SC，新聞 SC，CHJ では名詞比率と助詞比率には相関がない。名詞と最も強い負の相関が

あるのは助動詞で，動詞とその他は中程度の負の相関がある。

表 2　名詞比率を説明変数，主要品詞を目的変数とした回帰直線の傾き・切片・R^2

	助詞			助動詞			その他			動詞		
	傾き	切片	R^2	傾き	切片	R^2	傾き	切片	R^2	傾き	切片	R^2
図書SC	.015	.337	.001	−.488	.286	.520	−.294	.178	.301	−.232	.199	.285
新聞SC	−.004	.350	.000	−.450	.273	.507	−.348	.182	.451	−.198	.195	.299
雑誌SC	−.171	.390	.114	−.461	.280	.552	−.141	.134	.094	−.228	.197	.312
白書SC	−.185	.415	.084	−.343	.224	.297	−.268	.174	.239	−.204	.186	.194
CHJ	−.002	.340	.000	−.341	.220	.482	−.381	.183	.409	−.275	.258	.402

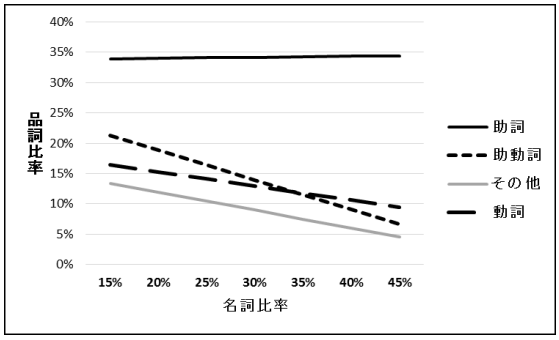

図 10 図書館 SC の回帰直線

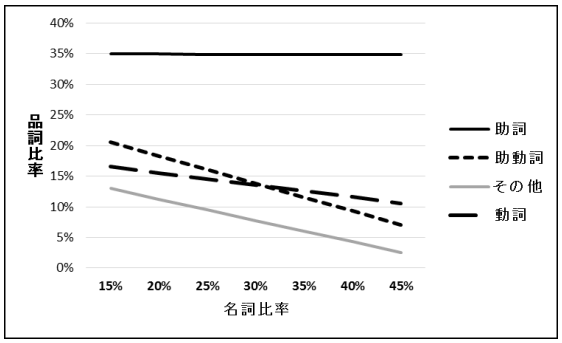

図 11 新聞の回帰直線

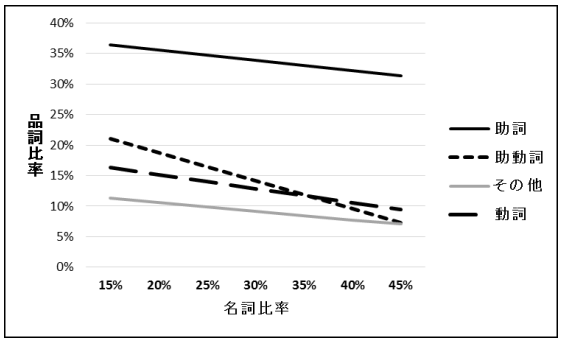

図 12 雑誌 SC の回帰直線

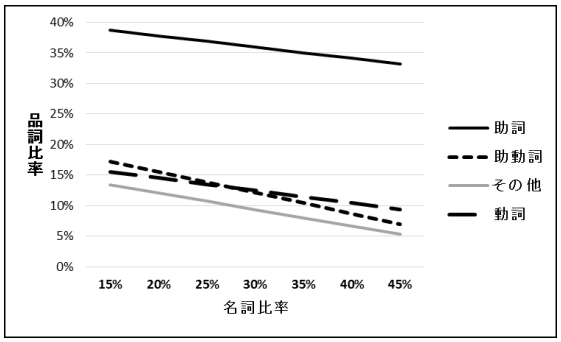

図 13 白書の回帰直線

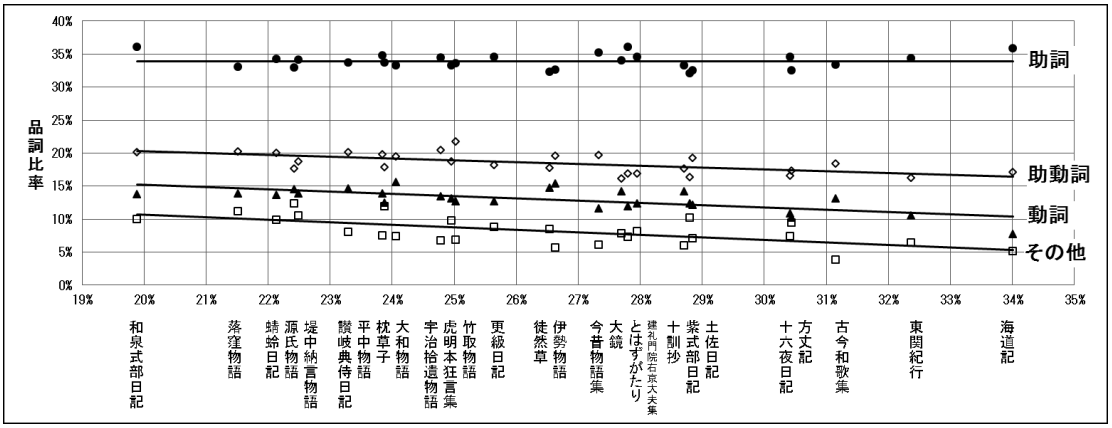

図 14 CHJ の回帰直線

図 14 の作品の並び順は，図 2 の大野（1956）の源氏物語，竹取物語，讃岐典日記，紫式部日記，土佐日記，枕草子，方丈記，徒然草，万葉集という順番とは若干異なっている。

４．考察①：名詞と助詞の大分類との相関

図 5～9 では，名詞比率が高くなると動詞比率・その他比率・助動詞比率が低くなる一方で，助詞比率はほぼ一定で変わらないように見える。ただし，図 10～14 の回帰直線では，図書 SC，新聞 SC，CHJ の助詞比率が X 軸と並行で名詞と無相関であるのに対し，雑誌 SC（R^2=.114）や白書 SC（R^2=.084）は緩やかな傾きがあり，弱い負の相関が見られる。

雑誌で名詞比率と助詞比率に弱い相関があるのは，名詞比率 45%未満のサンプルでも商品名や値段の列挙などが混入するサンプルが多いためだと考えられる。（6）は名詞比率 39.3%のサンプルである。

（6）キラキラとゴージャスなストーンがついたピアスたち。女の子らしくてちょっとよそ行きで、しぐさまでやわらかくなってくる。グリーン×パープル￥千二百　パープル×クリア￥千　ピンク￥千♥１♥２（BCCWJ サンプル ID：PM11_01212，実著者不明，『Ｍｙ　Ｂｉｒｔｈｄａｙ』，名詞比率 39.3%，助詞比率 25.6%）

雑誌 SC の場合，名詞比率が 40%程度でも商品の値段等が列挙されるサンプルが存在するため，名詞比率 45%未満・その他比率 30%未満という定義では，本研究で観察したい「一般的な日本語テキスト」には絞り込めていない可能性が高い。

白書は雑誌よりさらに名詞が列挙されるサンプルが多く，名詞比率 45%未満のサンプルは全体の 76.5%に留まる。白書も雑誌と同じように文書の絞り込みが十分にできていないため，弱い相関があると思われる。（7）は名詞比率 36.0%のテキストだが，名詞が列挙されている。

（7）消防関係者について，現在国が行っている表彰等には，日本国憲法に基づく栄典としての叙位，叙勲及び褒章，閣議決定に基づく内閣総理大臣表彰，消防表彰規程に基づく消防庁長官表彰並びに退職消防団員報償規程に基づく報償がある。これらの表彰等は，消防吏員，消防団員，消防教育職員及び消防機関並びに消防作業に協力した個人及び団体を対象として行われている。（BCCWJ サンプル ID：OW3X_00194，『消防白書』，昭和 63 年版，名詞比率 36.0%，助詞比率 29.1%）

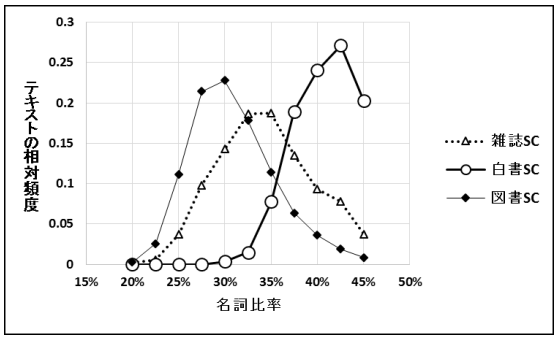

図 15　名詞比率の相対度数折れ線

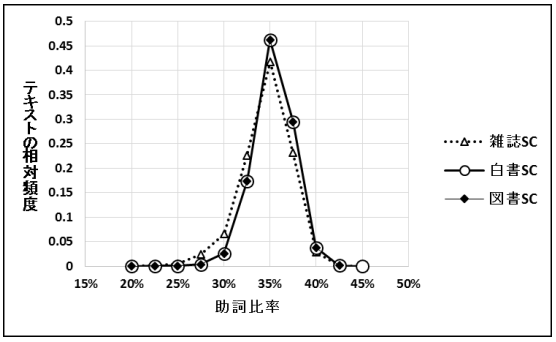

図 16　助詞比率の相対度数折れ線

図 15 は，横軸の名詞比率に対して図書 SC，雑誌 SC，白書 SC のテキストがどれぐらいの割合で存在しているかの分布を描いた図である。図書 SC＜雑誌 SC＜白書 SC の順に名詞の比率が高いテキストが多く分布しており，特に白書 SC は名詞比率の高いテキストが多いことが分かる。一方，図 16 は横軸を助詞比率にして，同様の分布を描いた図で，助詞比率は 3 つの SC で分布がほぼ同じであることが分かる。各 SC で助詞の分布がほぼ同じなのに，名詞の分布が異なるということは，雑誌 SC や白書 SC ではそれだけ助詞と結びつかない名詞の列挙が多いことを示唆している。雑誌 SC や白書 SC に名詞と助詞に弱い相関があるのは，名詞の列挙が多いテキスト等を除き切れていないことが原因である可能性が高い。図 16 の助詞比率はほぼ同じ分布をしていることから，雑誌 SC や白書 SC でもいわゆる一般的なテキストに絞り込めれば，名詞比率と助詞比率の相関はなくなると思われる。

名詞比率と助詞比率に相関が見られない一方で，もう一方の付属語である助動詞比率は最も名詞との負の相関が高いという正反対の結果となった。助動詞は動詞に接続する単語が多いため，名詞より動詞との相関が高そうに思われるが，助動詞と動詞の R^2 は図書 SC＝.048，新聞 SC＝.096，雑誌 SC＝.052，白書 SC＝.003，CHJ＝.176 と，CHJ を除けばほとんど相関はない。名詞比率が低いテキストには会話が多く含まれたり，難易度の低いテキストが多いことから，動詞の比率に連動しているというよりは，そのようなテキストに助動詞が使われやすく，文の凝縮度が高いテキストには助動詞が使われにくいことが考えられる。

本研究の目的の一つは，名詞と付属語の関係を明らかにすることにあった。一般的な日本語テキストでは，名詞比率と助詞比率に相関はないと考えられる。一方，名詞比率と助動詞比率は他の品詞より強い負の相関がある。

５．分析結果②：名詞比率と助詞中分類との相関

５.１　助詞中分類の代表的な単語

これまでは助詞の大分類による規則性を中心に観察してきた。本節からは助詞の中分類である格助詞・終助詞・係助詞・副助詞・準体助詞・接続助詞・連体助詞と名詞比率との相関を観察する。初めに図書 SC の中に出現した助詞の使用率 5 位までの例を示す（表 3)。

表 3　図書 SC10,551 テキストの中に出現した主な助詞とその使用率

格助詞		係助詞		接続助詞		副助詞		終助詞		連体助詞	
を	24.3%	は	73.9%	て	53.7%	か	20.6%	か	36.4%	の	100%
に	24.0%	も	25.5%	が	12.4%	や	17.1%	よ	18.4%	頻度	334,181
が	17.9%	こそ	0.5%	と	9.2%	など	11.1%	ね	16.3%		
と	13.4%	といっても	0.1%	ば	6.8%	まで	11.1%	な	10.5%	準体助詞	
で	8.5%	ぞ	0.02%	から	6.3%	だけ	8.9%	わ	4.7%	の	100%
頻度	1,198,605	頻度	295,021	頻度	182,690	頻度	93,158	頻度	54,816	頻度	32,028

助詞の頻度で見ると圧倒的に格助詞が多い。その格助詞も上位 5 種類だけで 88.1%となり，限られた助詞が多用されていることが分かる。本研究では UniDic で格助詞に分類されているノを連体助詞として独立した分類で扱っているが，格助詞の次に多いのがこの連体助詞である。3 番目に多いのは係助詞で，このほとんどはハとモである。接続助詞はテが最も多く，接続助詞全体の半分以上を占めている。副助詞は上位 5 種類で 68.8%になる。終助詞や準体助詞は頻度そのものが少なく，それぞれ格助詞の 4.6%，2.7%しかない。

５.２　名詞比率と助詞中分類との相関

次に第 3 節で行った助詞の大分類の観察と同様の方法で，今度は名詞比率と助詞の中分類の比率の関係を観察してみる。図 15～21 は助詞の中分類比率の積み上げ棒グラフである。

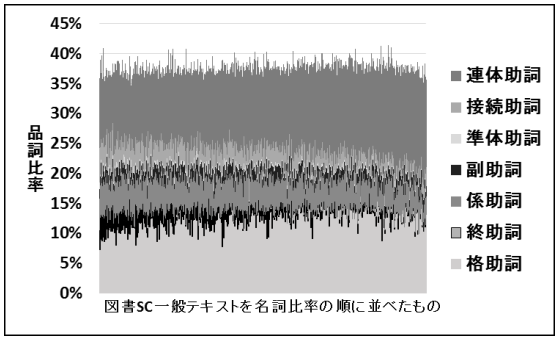

図 17 図書 SC の助詞比率，N=10,364

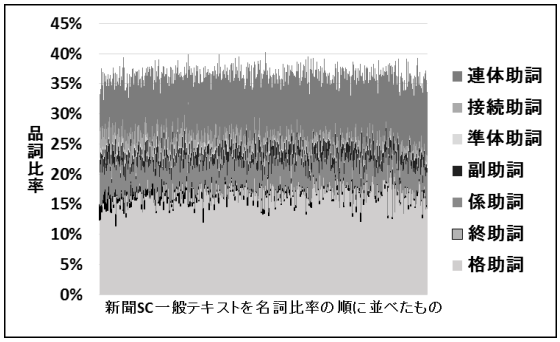

図 18 新聞 SC の助詞比率，N=1,473

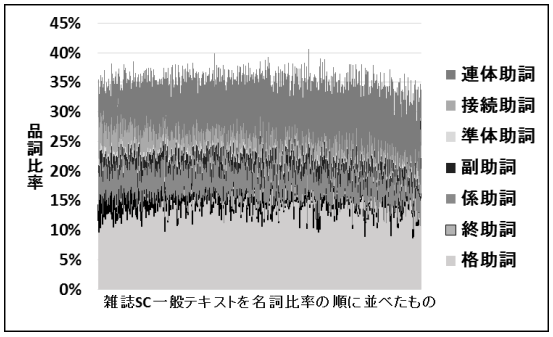

図 19 雑誌 SC の助詞比率，N=1,690

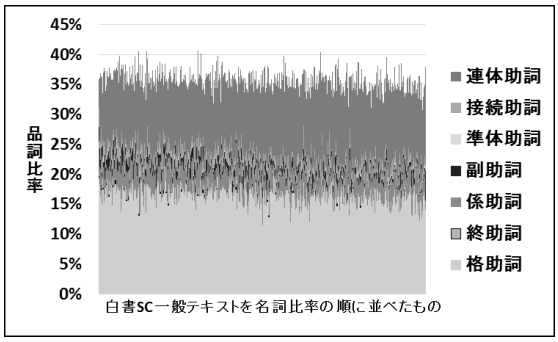

図 20 白書 SC の助詞比率，N=1,147

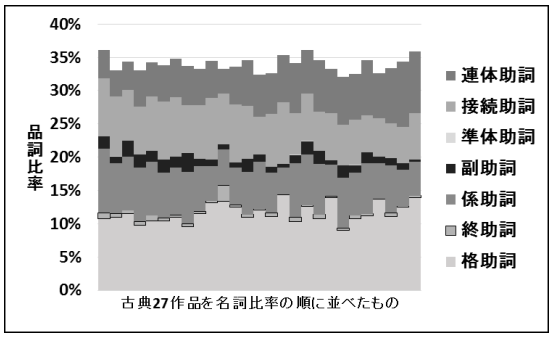

図 21 CHJ27 作品の助詞比率，N=27

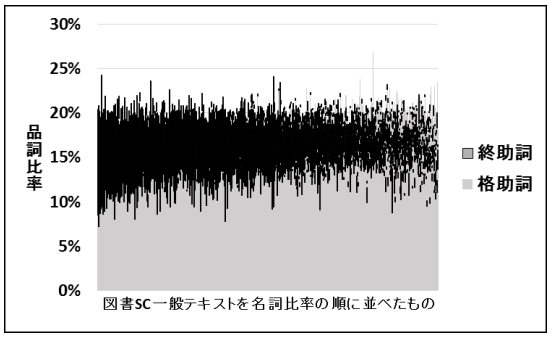

図 22 図書 SC の格助詞と終助詞，N=10,364

これらを見ると，助詞の中分類には名詞比率に相関して増減している種類があることが分かる。多くのコーパスに見られる規則性は名詞比率に対する連体助詞比率の正の相関と，接続助詞比率の負の相関で，この 2 つを合計するとほぼ一定の比率になるように見える。図 22 は図 17 から格助詞と終助詞を抜き出して拡大して描画した図である。図 22 では終助詞がやや強調されて描かれているが，興味深いことに格助詞と終助詞を加えるとその割合はほぼ一定になるように見える。品詞の大分類で観察したときは，名詞比率と助詞比率にはほとんど相関がなかった。しかし，助詞の機能で分けた中分類では様々な相関が観察される。注目されるのは個別の中分類では名詞比率との相関がありながら，それらの一部や全部を

合計するとほぼ一定となるという規則性である。

次に名詞を説明変数，助詞の中分類を目的変数とした回帰分析の結果を示す。表 4 では助詞の中分類で比率の高い格助詞・係助詞・接続助詞・連体助詞のみを示した。図 23～27 はこれを図示したものである。

表 4 名詞比率を説明変数，主要助詞を目的変数とした回帰直線の傾き・切片・R^2

	格助詞			係助詞			接続助詞			連体助詞		
	傾き	切片	R^2	傾き	切片	R^2	傾き	切片	R^2	傾き	切片	R^2
図書SC	.165	.110	.134	−.055	.070	.038	−.155	.079	.303	.285	−.230	.428
新聞SC	.116	.138	.055	−.072	.078	.055	−.123	.065	.265	.150	.017	.148
雑誌SC	.023	.152	.003	−.052	.070	.037	−.130	.072	.289	.130	.072	.155
白書SC	−.139	.234	.053	−.049	.058	.010	−.054	.039	.052	.146	.039	.046
CHJ	.156	.075	.144	−.185	.116	.243	−.249	.139	.472	.392	−.040	.809

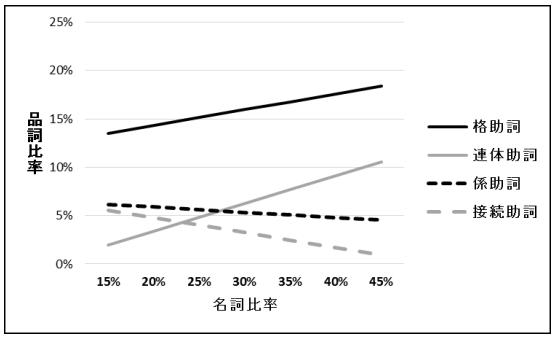

図 23 図書館 SC の回帰直線

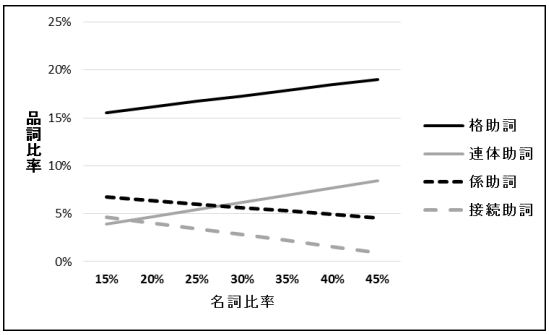

図 24 新聞の回帰直線

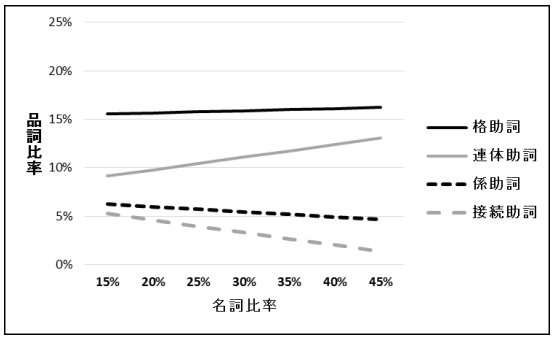

図 25 雑誌 SC の回帰直線

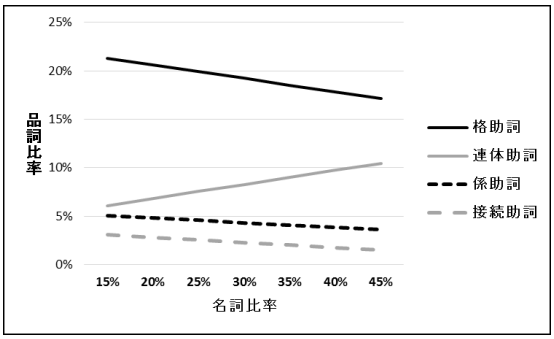

図 26 白書の回帰直線

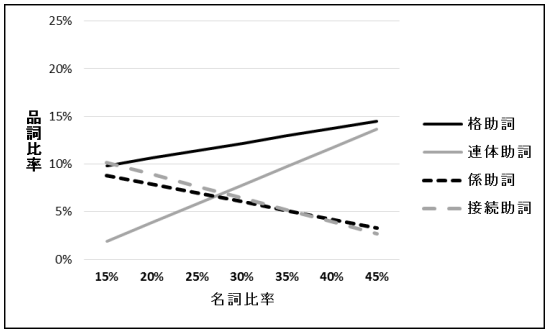

図 27 CHJ の回帰直線

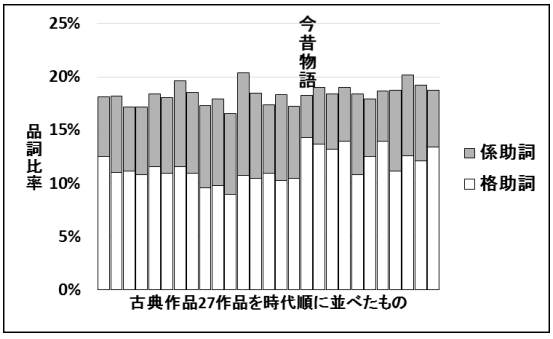

図 28 CHJ の格助詞と係助詞の比率・時代順

品詞の大分類での分析と同じように図書 SC，新聞 SC，CHJ は似た傾向を示し，雑誌 SC と白書 SC はやや異なる傾向となっている。ただし，連体助詞（荒い鎖点）が上向きで接続助詞（グレーの線）が下向きになるという傾向は全てに共通しており，表 4 の決定係数 R^2 を見ても白書を除くといずれも比較的高い相関が確認できる。また図書 SC，新聞 SC,CHJ では格助詞と係助詞に正反対の相関が観察される。

なお，図 26 は CHJ の作品を時代順に並べた時の格助詞と係助詞の比率の推移である。図の 2/3 付近で格助詞比率が急に高くなるのが今昔物語でこの作品から右側が鎌倉時代の作品である。鎌倉時代以降，格助詞の比率が高くなっている作品が多い。

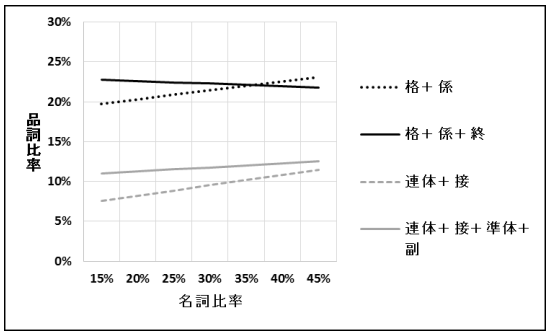

図 29 図書館 SC の回帰直線

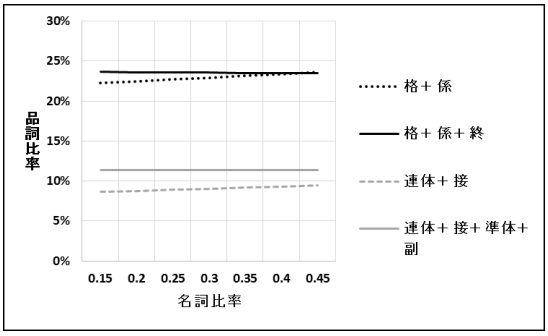

図 30 新聞の回帰直線

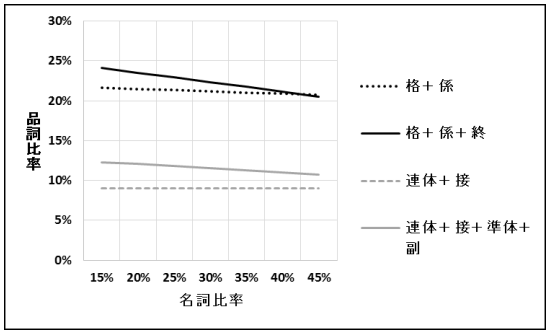

図 31 雑誌 SC の回帰直線

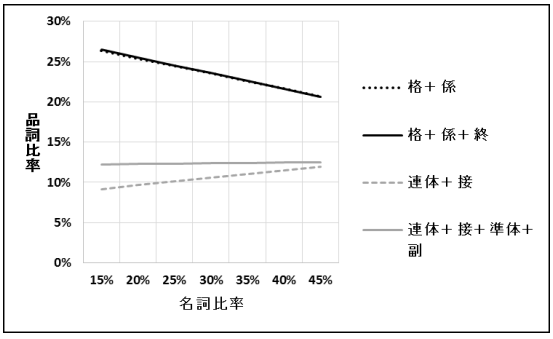

図 32 白書の回帰直線

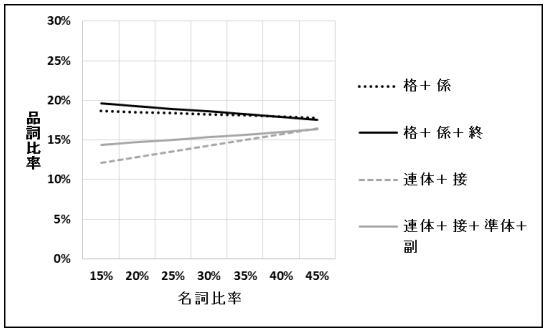

図 33 CHJ の回帰直線

表5 図29～33の決定係数

	格＋係	格＋係＋終	連体＋接	連体＋接＋準体＋副
図書SC	.056	.007	.118	.015
新聞SC	.008	.000	.008	.000
雑誌SC	.004	.075	.000	.023
白書SC	.071	.075	.019	.000
CHJ	.013	.053	.175	.041

図 23～27 では連体助詞と接続助詞，格助詞と係助詞が名詞に対して正反対の相関になっている傾向が見られた。第 3 節の分析では助詞の大分類と名詞との相関は観察されなかったことから，格助詞と係助詞，連体助詞と接続助詞を足し合わせると名詞との相関がなくな

ることが予想される。そこで図 29〜33 では名詞と格助詞＋係助詞と連体助詞＋接続助詞を小計した比率との回帰直線を観察する。また図 22 では，格助詞と終助詞を加えると比率が一定になる傾向が見られたことから，格助詞＋係助詞＋終助詞と残りの連体助詞＋接続助詞＋準体助詞＋副助詞の回帰直線も一緒に描くことにする。

図 29〜33 では，図 23〜27 の黒の線同士を合計したものを黒の点線，灰色の線同士を合計したものを灰色の点線で表示している。黒の実線はさらに終助詞を加えたもので，灰色の実線は準体助詞と副助詞を加えたものである。図 23〜27 では正と負に相関があった助詞同士を合計すると相関が低くなる傾向が観察される。特に図書 SC と新聞 SC では格助詞＋係助詞＋終助詞のグループと，残りの連体助詞＋接続助詞＋準体助詞＋副助詞のグループの小計でほぼ名詞との相関がなくなる。

雑誌 SC と白書 SC の格助詞＋係助詞＋終助詞が負の相関になるのは，図 12，13 の助詞の大分類の傾向と同じである。図 32 の白書 SC で線が 3 本になっているのは，白書 SC ではほとんど終助詞が出現しないため，格助詞＋係助詞の合計と格助詞＋係助詞＋終助詞の合計がほぼ同じになるためである。

６．考察②：名詞比率と助詞中分類との相関

名詞比率と助詞の大分類に相関はなかった。しかし，助詞の機能で分けた中分類では様々な相関が見られた。基本的に格助詞と係助詞，連体助詞と接続助詞は正反対の相関がある。これらを加えると名詞との相関が小さくなり，格助詞＋係助詞＋終助詞のグループと，残りの連体助詞＋接続助詞＋準体助詞＋副助詞のグループの小計でほぼ名詞との相関がなくなる。このことは一般的な日本語テキストにおいて名詞の比率と助詞の大分類に相関がない理由として，格関係を表す働きと，接続関係を表す働きを中心とした助詞の働きが別々に関係していることを示唆している。

格助詞＋係助詞＋終助詞の合計がほぼ水平になるのは図書 SC と新聞 SC のみで雑誌 SC と白書 SC は大分類と同じように右肩下がりになる。この理由は第 4 節で考察した通り，「一般的なテキスト」の絞り込みがうまくいっていないためだと考えられる。一方，大分類では水平だった CHJ で傾きが生じるのは，格助詞と連体助詞の分類の仕方が荒かったためだと思われる。古典語の場合，「君が代」「松が枝」のように現代語の連体助詞ノが使用される場所で格助詞のガが使用されることがある。このようなガは格助詞ではなく，連体助詞として認定する必要があり，これを分類し直すと格助詞の比率は下がり，連体助詞の比率は上がって傾きが水平に近くなる可能性がある。

図 28 は時代順に格助詞と係助詞の比率を並べたグラフで，時代が新しくなるにつれ係助詞率が減り，格助詞率が増える様子が見られる。この図は係り結びが衰退し格助詞が発達する歴史的変化を表していると考えられる。興味深いのは係助詞と格助詞を足すとほぼ一定の割合になることで，格関係を表す助詞は 2 割弱と一定で，その範囲内で係助詞と格助詞の交代現象が起こったように見える。

連体助詞は名詞と名詞をつなぐ助詞であるから名詞が増えると連体助詞が増えるのは理解しやすい。その逆に名詞が増えれば動詞や助動詞は減少するため，これらに接続する接続助詞が減少するという関係も納得できる。接続助詞と動詞の決定係数 R^2 は図書 SC＝.297，新聞 SC＝.110，雑誌 SC＝.251，白書 SC＝.034，CHJ＝.670 と，白書 SC を除けば中程度〜強い相関がある。

しかし 2 つのグループは格関係と接続関係を表す働きで完全に二分されるわけではなく，

なぜこのようなグループだと相関がなくなるのかは，現段階ではよく分からない。終助詞が出現するテキストは会話などの口語的な文体のテキストが含まれている場合がほとんどで，これらの場合，必須格が無助詞化することが多い。文単位で終助詞の数と無助詞の数が一致するわけではないが，終助詞が多用されるテキストほど格助詞が省略される回数が多くなり，トータルで見ると終助詞の数と無助詞の数がほぼ同じになっていると思われる。このため，終助詞が格関係と全く無関係であるわけではない。

だが，機能的に格助詞のグループに近い副助詞は，格助詞のグループではなく，連体助詞のグループに入れたほうが 2 つのグループ小計の回帰直線の傾きが水平に近くなる。また，準体助詞は名詞や節を接続する機能は持たないが，これも連体助詞のグループに加えたほうがグループ小計の回帰直線の傾きが水平に近くなる。確かに連体助詞と接続助詞を足すと名詞との相関は小さくなるが，決定係数が新聞 SC：.008，雑誌 SC：.000，白書 SC：019 で小さい一方で，図書：.118，CHJ：.175 では一定の相関が存在する。図書と CHJ ではこれにさらに準体助詞と副助詞の比率を加えることで図書：.015，CHJ：.041 のように相関がなくなるのである。この理由については品詞の分類法やグループ分けの方法を含め，さらに調査していく必要がある。

５．まとめと今後の課題

日本語のテキストでは名詞比率に連動して動詞や形容詞などの比率が規則的に変化することが知られている。しかし名詞比率と付属語の関係は明らかにされていないため，本研究では BCCWJ 固定長・長単位データと CHJ 長単位データを使用し，名詞比率と助詞比率の相関を中心に観察を行った。観察に当たっては名詞の列挙やカタカナ文の形態素解析ミスを多く含んだテキストを対象から除くため，「名詞比率 45%未満・その他比率 30%未満」のサンプルを仮に「一般的な日本語テキスト」と定義し，これを分析に使用した。観察の結果，助動詞比率は名詞と強い相関があるのに対し，助詞比率の大分類にはほとんど相関がなく，名詞の比率に関わらず，多くのテキストで 34%前後の比率で使用されていることが分かった。

しかし，助詞の機能別に分類した中分類では，格助詞と連体助詞などには正の相関が，係助詞と接続助詞などには負の相関があるなど，助詞の機能によって様々な相関が観察された。その一方で格関係の働きを中心とした格助詞＋係助詞＋終助詞のグループと，残りの連体助詞＋接続助詞＋準体助詞＋副助詞のグループの小計では名詞との相関がほとんどなくなることが分かった。助詞の大分類で相関がなくなる理由には，格関係を表す働きと，接続関係を表す働きを中心とした助詞の働きが別々に関係していることが考えられる。

本研究に残された課題は多い。第 1 に日本語のテキストから一部のテキストだけ抜き出して観察する是非を含めて，「一般的な日本語テキスト」の定義とその抽出方法を精密化していく必要がある。第 2 には品詞の分類方法をもっと精密化させていく必要がある。特に古典語の場合，格助詞と連体助詞を区分するためには助詞の使われ方を観察して分類しないと正確な分析は難しい。これらは本研究を精密化させるために必要な課題である。

第 3 の課題として，もっと本質的な問いである「助詞の比率が一定であることの意義」を解明していく必要がある。他の単語が名詞の比率に対してすべて相関があり，助詞も中分類では相関があるにも関わらず，助詞の大分類ではなぜ相関がなくなるのかの解明は，興味深い課題である。この解明に向けて，膠着語ではいわゆる膠の働きをする単語がどの言語でも一定の比率になっているのか，他の言語を調査することも必要である。また，現段階ではこの現象が文法に関わる現象なのか，文体に関わる現象なのかも判然としないが，もしこれが文法の問題であるとするなら，日本語を大きく自立語と付属語に分け，付属語を助詞と助動詞に分けるといった品詞体系を見直す切り口になるかも知れない。

第 4 には，なぜ助詞の比率が 34%前後で一定になるのかを解明する必要がある。助詞は単語と単語の間で使用されるため，日本語が「単語＋助詞＋単語」という基本構造を持つのであれば，その 1/3 に当たる 34%前後で一定になるのも当然かもしれない。しかし，日本語には 1 単語で 1 文として完結する文もあり，すべての文が「単語＋助詞＋単語」の構造を持つわけではない。助詞比率がなぜ 34%前後で一定になるのかを解明するには，そもそもどんな分析方法を使用すればそれが解明できるのかといったごく根本的なところから考察していく必要がある。

このような解明は，個人による分析では到達解決困難だと思われる。今後，多くの研究者が本研究に関心を持ち，様々な角度から解明が進むことを期待したい。

使用データ

国立国語研究所のプロジェクトによる成果『現代日本語書き言葉均衡コーパス』ならびに『日本語歴史コーパス』および検索システム・コーパス検索アプリケーション「中納言」（バージョン 2.2.3.1：https://chunagon.ninjal.ac.jp/ijas/search）を利用して行われたものである。

〈参考文献〉

大野晋（1956）「基本語彙に関する二三の研究」『国語学』24，pp.34-46.

樺島忠夫（1955）「類別した品詞の比率に見られる規則性」『国語国文』24（6），pp.385-387.

柏野和佳子（2013）「書籍サンプルの文体を分類する」『国語研プロジェクトレビュー』Vol.4 No.1：43-53.

国立国語研究所（2015）『BCCWJ 図書館サブコーパスの文体情報』（第 1 版）.
データは http：//pj.ninjal.ac.jp/corpus_center/anno/からダウンロードできる.

冨士池優美・小西光・小椋秀樹・小木曽智信ほか（2011）「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」『言語処理学会第 17 回年次大会発表論文集』，pp.663-666.

山崎誠（2014）「言語単位と文の長さが品詞比率に与える影響」『第 5 回コーパス日本語学ワークショップ予稿集』国立国語研究所，pp.233-242.