Relationship between Literary Style and Modifying Particle "no"

連体助詞の「ノ」と文体の関係

森　秀明（東北大学文学研究科）

Relationship between Literary Style and Modifying Particle "no"

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

要旨

名詞の頻度と文体には強い関連性があり，硬い文体や客観的な文体ほど名詞の頻度が高い。一方，連体助詞の「ノ」は名詞の頻度に連動して増減する。それでは，ノと文体の関係はどうだろう。硬い文体は難易度が高い傾向があるため，名詞の増加以上にノが増加するのだろうか。本研究では BCCWJ 図書館書籍に文体指標をつけた国立国語研究所（2015）を利用し，文体の違いによる名詞とノの回帰直線を調査した。回帰直線は外れ値に弱く，この除去が分析のカギとなる。図書館書籍では固有名詞や数詞が列挙されるサンプルが存在し，これらが外れ値となっている。そこで文書構造タグの<figureBlock>と<list>が存在するサンプルを除き，普通名詞と普通名詞に接続するノに絞って回帰直線を調査した。その結果，文体による変化は見られなかった。ノは，普通名詞に連動して増減するだけで，その頻度に文体の影響はない。ノの頻度は，人間の意志や個性とはほとんど無関係に増減している可能性がある。

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品詞構成比率（図 2）を分析し，これを明らかにした。

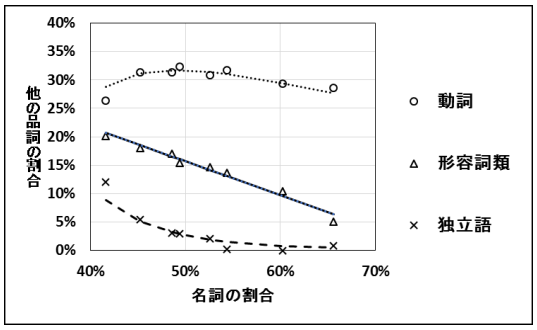

図 1：樺島（1955）第一表に基づく散布図

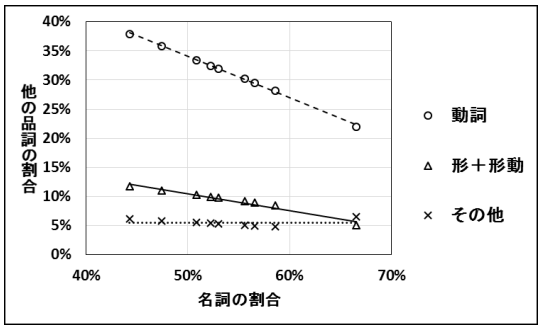

図 2：大野（1956）第七表に基づく散布図

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では名詞と付属語の関係は不明なままであったため，発表者は BCCWJ・固定長・長単位データを使用してこれを調査し，昨年度の言語資源活用ワークショップで報告した（森，2017）。品詞比率を調査する際，図鑑などのサンプルでは名詞が多数列挙され，もはや文章とは呼べ

ないテキストも存在したため，名詞比率 45%未満等の条件で絞り込んだサンプルを仮に「一般的な日本語テキスト」と定義してこれを分析に使用した。主な調査結果は次の通りである。

(1) 助動詞と名詞には，強い相関がある。
(2) 助詞と名詞には，相関関係がない（日本語のテキストで助詞比率はほぼ一定である）。
(3) 連体助詞と名詞，接続助詞と名詞には中程度の正と負の相関関係があるが，これらを「語と語を結合する助詞」と考えて頻度を合計すると，この合計数と名詞との相関が低くなる（日本語のテキストで結合助詞比率はほぼ一定である）[1]。
(4) 格助詞や係助詞などと名詞には正と負の相関があるが，結合助詞以外の助詞の全てを「格関係に関わる助詞」と考えて頻度を合計すると，この合計数と名詞との相関が低くなる（日本語のテキストで格関係助詞比率はほぼ一定である）。

表 1　品詞頻度の積率相関行列：BCCWJ 図書館書籍 SC の一般文書　n =10,385

	名詞	普通名詞	動詞	その他	助動詞	助詞	結合助詞	格関係	格助詞	係助詞	終助詞	準体助詞	副助詞	接続助詞	連体助詞
普通名詞	0.814														
動詞	−0.533	−0.423													
その他	−0.545	−0.410	−0.003												
助動詞	−0.722	−0.668	0.222	0.231											
助詞	0.024	0.089	−0.106	−0.326	−0.381										
結合助詞	0.340	0.340	−0.250	−0.259	−0.426	0.429									
格関係助詞	−0.261	−0.195	0.103	−0.112	−0.028	0.652	−0.405								
格助詞	0.362	0.415	−0.007	−0.480	−0.419	0.400	0.036	0.374							
係助詞	−0.198	−0.222	−0.118	0.087	0.150	0.223	−0.190	0.385	−0.249						
終助詞	−0.517	−0.580	0.230	0.321	0.365	−0.003	−0.325	0.270	−0.514	0.005					
準体助詞	−0.441	−0.428	0.133	0.208	0.286	0.171	−0.245	0.379	−0.238	0.119	0.425				
副助詞	−0.193	−0.045	0.028	0.157	0.019	0.184	−0.159	0.320	−0.227	−0.030	0.159	0.145			
接続助詞	−0.549	−0.486	0.546	0.199	0.199	0.156	0.155	0.027	−0.285	0.007	0.349	0.225	0.098		
連体助詞	0.650	0.610	−0.570	−0.355	−0.499	0.274	0.771	−0.370	0.215	−0.170	−0.508	−0.359	−0.202	−0.509	
普通名詞＋ノ	0.579	0.708	−0.456	−0.305	−0.472	0.223	0.686	−0.350	0.196	−0.193	−0.477	−0.342	−0.132	−0.434	0.877

表 2　品詞頻度の積率相関行列：BCCWJ 新聞 SC の一般文書　n =1,353

	名詞	普通名詞	動詞	その他	助動詞	助詞	結合助詞	格関係	格助詞	係助詞	終助詞	準体助詞	副助詞	接続助詞	連体助詞
普通名詞	0.631														
動詞	−0.560	−0.346													
その他	−0.673	−0.470	0.110												
助動詞	−0.714	−0.664	0.318	0.456											
助詞	−0.028	0.278	−0.098	−0.304	−0.486										
結合助詞	0.050	0.215	−0.228	−0.053	−0.218	0.373									
格関係助詞	−0.063	0.139	0.055	−0.277	−0.351	0.774	−0.299								
格助詞	0.228	0.252	0.097	−0.534	−0.425	0.526	−0.168	0.655							
係助詞	−0.442	−0.407	0.150	0.432	0.382	−0.132	−0.083	−0.079	−0.396						
終助詞	−0.232	−0.182	−0.094	0.230	0.068	0.204	−0.160	0.319	−0.277	0.056					
準体助詞	−0.422	−0.296	0.127	0.338	0.264	0.078	−0.100	0.148	−0.199	0.358	0.175				
副助詞	0.053	0.304	−0.073	−0.118	−0.237	0.347	−0.018	0.369	−0.028	−0.019	−0.122	−0.026			
接続助詞	−0.517	−0.457	0.340	0.451	0.386	−0.142	0.172	−0.264	−0.424	0.403	0.108	0.232	−0.105		
連体助詞	0.363	0.475	−0.415	−0.325	−0.434	0.423	0.796	−0.108	0.109	−0.323	−0.210	−0.233	0.048	−0.460	
普通名詞＋ノ	0.215	0.566	−0.260	−0.168	−0.335	0.330	0.675	−0.121	−0.018	−0.199	−0.154	−0.131	0.111	−0.299	0.792

このことから，日本語は文体や叙述内容に関わらず，文章を書く際には全語数の約 1/3 を

[1] 表 1 の図書館書籍 SC の場合、名詞と結合助詞の相関は.340 で弱い相関があるため、「日本語のテキストで結合助詞比率はほぼ一定である」とは言えないが、表 2 の新聞 SC では.050 と相関がないため、仮にこのように考える。格関係助詞も同じである。

助詞に使用し，助詞の 1/3 は語と語の結合に，残りの 2/3 は格関係に使用するシステムを持っていると考えられる。

この発表に対しては大きくまとめて次の 2 点の指摘を頂戴した。

（5）これだけでは，この現象がどのような言語学的意味を持っているのか分からない。
（6）全データを使用していない分析結果を，日本語全体に一般化することはできないのではないか。

本研究ではこれを受けて，この問題をさらに検討する。

日本語における品詞比率の問題は，文体との関連で論じられてきた。樺島（1955：386）では「名詞の百分率をもって，文章の特性を計る尺度となし得る」とされ，「N の増加は話し言葉的なものから書きことばへと向かつている」，「感情の表現をなすものから關係の表現をなすものへと，N が増す」（p.387）などの特徴が指摘されている。

しかし，どのような文章においても助詞の比率が一定であるということは，助詞はそのような文体や叙述内容とは無関係に使用されていることを示唆している。ただし，助詞全体の比率は一定であったが，例えば連体助詞には名詞と正の相関が，接続助詞には負の相関があった。名詞に対して相関があるという点では動詞や助動詞などと同じである。このため，対象を連体助詞ノに絞り，文体の違いによってノがどのように使用されているかを調査することにする。

BCCWJ 図書館書籍サブコーパス（以下図書 SC と略す）には，10,551 文書を人手で判断して文体情報を付与した国立国語研究所 (2015)『BCCWJ 図書館サブコーパスの文体情報』[2]が存在し，その詳細は柏野（2013）で紹介されている。ここでは図書 SC のサンプルをテキスト構造が単純なもの（例：章 節構造）と，テキスト構造・紙面形式などの点で文体の評定値をつけるのになじまないもの（全体の約 2 割）に分け，前者には「専門度，客観度，硬度，くだけ度，語りかけ性度」といった評定値が，後者には「対談，Q&A 形式，図解，用語解説」等の分類情報が付与されている。本研究ではこの文体指標を利用し，文体の違いによってノの使用に変化があるかどうかを調査する。

硬い文体と軟らかい文体，客観的な文体と主観的な文体などでは，名詞の頻度が異なることが知られている。硬い文体や客観的な文体ほど名詞の頻度が高く，凝縮的な文体になっている（樺島，1955）。その一方で，連体助詞ノは，名詞の頻度に連動して増減することが知られている（森，2017）。それでは，ノと文体の関係はどうであろうか。硬い文体や客観的な文体では名詞の頻度が高いため，ノの頻度が高くなるのは当然だが，これらの文体のテキストは複雑で難易度の高い内容を記述していることが多いことから，名詞が多くなった以上にノの頻度が高くなるのであろうか。

これを調査するには，サンプルごとの名詞の頻度を X 軸（説明変数），ノの頻度を Y 軸（目的変数）とする散布図に回帰直線を描き入れて，この傾きや切片が硬いテキストと軟らかいテキストによって異なるのかどうかを観察すればよい。回帰直線の傾きや切片が異なるなら，ノの頻度は文体によって使い分けられているし，これが同じであれば文体による使い分けはないと考えられる。

[2] http://pj.ninjal.ac.jp/corpus_center/anno/の「サンプルに対する文体指標（sty）」で，BCCWJ_LB_Stylistics-1.0.zip のファイルが公開されている。

ノの使用が文体とは無関係に行われているとすれば，日本語のテキストにおいて助詞の比率が一定である理由も分かりやすくなる。すなわち，助詞は文体や叙述内容に関わらず，語の結合と格関係の表示に一定数を必要とするシステムで，そこに人間の意志や個性はほとんど介在していない。日本語文法では助詞・助動詞を付属語とか機能語という扱いで同列に扱ってきたが，真に日本語の機能をつかさどっている品詞は助詞であり，助動詞は名詞に連動して増減する点において，動詞や形容詞と同じ分類に入る品詞である。このことは，助動詞が文体や叙述内容に深くかかわる品詞であることを示唆している。

（6）のデータ選択の問題は，コーパス言語学において古くから論じられてきた問題で，基本的には分析に適さないデータは除くべきだと考えられる。国立国語研究所（2015）でも文体の評定値をつけるのが難しかったサンプルが 2 割ほど存在することが指摘されている。これらは，「対談，Q&A 形式，図解，用語解説」等のサンプルで，文体分析に使いたくともその性質が文体評定に適さないため，評定がつけられなかったサンプルである。前回発表でも国立国語研究所（2015）に従って分析を行おうと試みたが，なお分析に適さないと考えられるサンプルが存在したため，名詞比率 45%未満という基準を設けた。しかし，このような一律に足切りをする基準では，恣意的なデータ選択の印象を免れないため，今回は，図書 SC を構築する際に付与された文書構造の情報に基づいてサンプルの選択を行い，データ選択の違いによって分析結果にどのような影響が出るのかについても考察する。

２．分析データ

２.１　使用するコーパスとデータの種類

分析には図書 SC の固定長・長単位データを使用する。BCCWJ 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，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[3]。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 3 は，図書 SC 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。本研究では何も絞り込みを行わないデータをフルデータと呼ぶ。

図 3 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 4 は国立国語研究所（2015）の文体情報を使用し，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 3 と同様に描いた

[3] 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。

散布図である。本研究ではこれを章節構造データと呼ぶ。

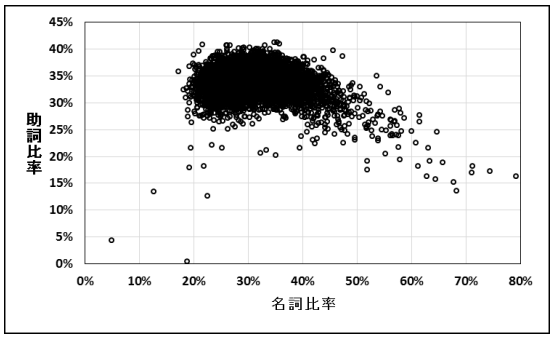

図 3 名詞比率と助詞比率の散布図：
図書 SC フルデータ，N=10,551

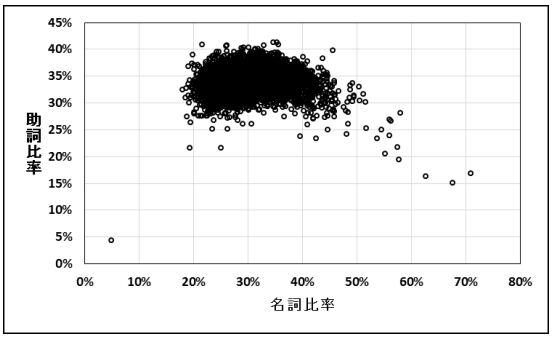

図 4 名詞比率と助詞比率の散布図：
図書 SC 章節構造データ，N=8,792

「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（7）は「視覚表現多用系」，（8）は「データベースやリスト系」の文書の一部である。

（7）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%）

（8）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%）

（7），（8）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 4 のように尾の部分がかなり少なくなる。それでもまだ図 4 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 4 の尾の部分のサンプルを観察する。（9）は図 4 で最も名詞比率が高いサンプル（10）は名詞比率 44.5%のサンプル（11）は名詞比率が最も少ないサンプルである。

（9）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%）

（10）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%）

（11）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%）

（9）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。（10）も後半は名詞の列挙で一般的な文章になっていない。(11) 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。本研究の目的は文体の違いによって名詞とノの回帰直線に変化があるかどうかを調査することにあるため，ノが出現する余地なく名詞が列挙されているサンプルを含めて分析する意義は低いと考えられる。

前回の発表では名詞の列挙を含む文を少なくする目的で名詞比率を 45%未満に絞り込み，解析ミスと考えられる「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% 未満に絞り込んだ。その上で，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した。しかし，このような絞り込みでは恣意的なデータ選択を行っている印象はぬぐえない。そこで本研究では，BCCWJ にタグ付けされている文書構造の情報を利用してサンプルの絞り込みを行う。

文書構造タグは，原資料を電子的なテキストに変換するに当たって，元々の資料が持っていた構造を復元できるように付与された情報である（詳しくは，山口，2014；西部・大島・間淵・小林ほか，2011；山口・高田・北村・間淵，2011 を参照のこと）。図 5 は図表からサンプルを取得する際につけられた文書構造タグの例である。上の段は原資料の表が，下の段はそれを電子化したデータが表示されている。

図 5 の左の表は，上段に表のキャプションがあり，下に項目と数字の表がある。下段では先頭に<figureBlock>, 最後に</figureBlock>というタグが付与されている。<figureBlock>とは，「図表・写真・絵などの要素と，それに付随する文書要素をまとめた要素を表す。」（山口・高田・北村・間淵，2011:82）。つまり<figureBlock>のタグが付いていると，その文書には図表・写真・絵などの要素が含まれていることを意味している。左の表は，数字が主体であるため，サンプリングされたのは表のキャプションのみである。

これに対し，右の図は言語情報が主体であるため，表の中のリストがデータとして採取されている。この時つけられているのが<list>というタグである。<list>は「箇条書きなど，列挙された文書要素の集まりを表す。」（山口・高田・北村・間淵，2011:91）。つまり<list>のタグが付いていると，その文書には名詞の列挙が含まれる可能性が高くなる。

〔PB24_00304：『生命倫理とこころのケア』〕

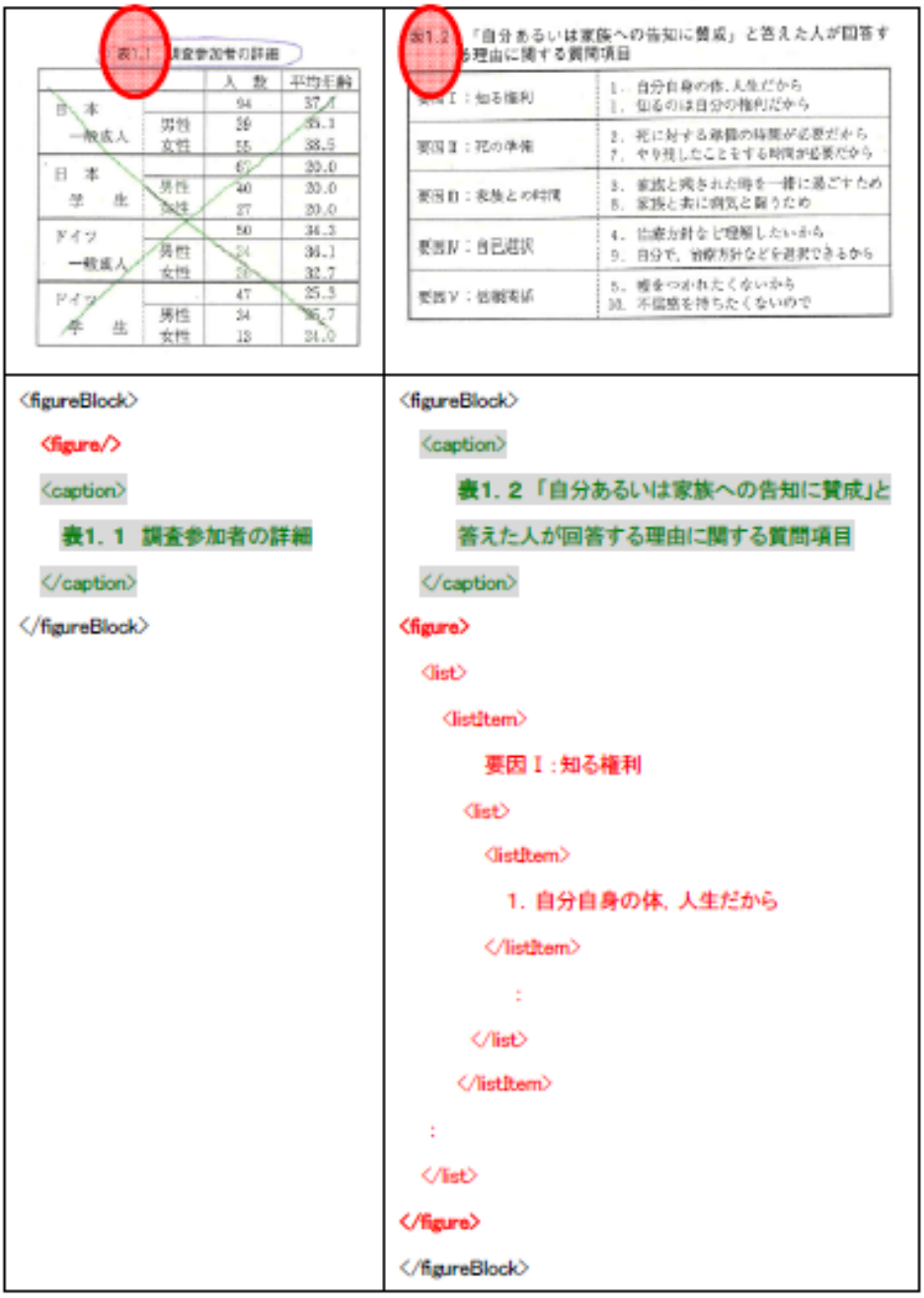

表1.1 調査参加者の詳細

		人数	平均年齢
日本 一般成人		94	37.4
	男性	39	35.1
	女性	55	38.5
日本 学生		67	20.0
	男性	40	20.0
	女性	27	20.0
ドイツ 一般成人		50	34.3
	男性	24	36.1
	女性	26	32.7
ドイツ 学生		47	25.3
	男性	34	26.7
	女性	13	21.0

表1.2 「自分あるいは家族への告知に賛成」と答えた人が回答する理由に関する質問項目

要因Ⅰ：知る権利	1．自分自身の体，人生だから 1．伝えるのは自分の権利だから
要因Ⅱ：死の準備	2．死に対する準備の時間が必要だから 7．やり残したことをする時間が必要だから
要因Ⅲ：家族との時間	3．家族と残された時を一緒に過ごすため 8．家族と共に病気と闘うため
要因Ⅳ：自己選択	4．治療方針など理解したいから 9．自分で，治療方針などを選択できるから
要因Ⅴ：情報獲得	5．嘘をつかれたくないから 10．不信感を持ちたくないので

```
<figureBlock>
  <figure/>
  <caption>
    表1．1　調査参加者の詳細
  </caption>
</figureBlock>
```

```
<figureBlock>
  <caption>
    表1．2「自分あるいは家族への告知に賛成」と
    答えた人が回答する理由に関する質問項目
  </caption>
<figure>
  <list>
    <listItem>
      要因Ⅰ：知る権利
      <list>
        <listItem>
          1．自分自身の体，人生だから
        </listItem>
          :
      </list>
    </listItem>
  :
  </list>
</figure>
</figureBlock>
```

図 5　図表からのサンプリングの例（西部・大島・間淵・小林ほか，2011:271 より引用）

本研究では，<figureBlock>と<list>のタグが含まれている文書は，名詞の頻度を使用した文体分析には適さない文書と判断し，これを除いた文書で分析を行う。本研究ではこれを選抜データと呼ぶ。この基準によって除かれる文書数は 2,266 文書で，残存率は 78.5%（8,283 文書）である。また選抜データと国立国語研究所（2015）の章節構造データの基準を同時に適用した際に除かれる文書数は 3,362 文書で，残存率は 68.1%（7,189 文書）である。本研究ではこれを二重選抜データと呼ぶ。

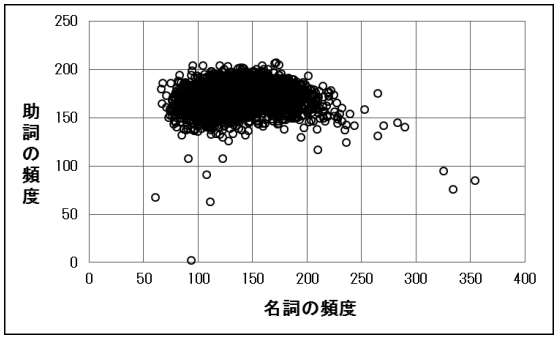

図 6 名詞頻度と助詞頻度の散布図：
図書 SC 選抜データ，：N=8,283

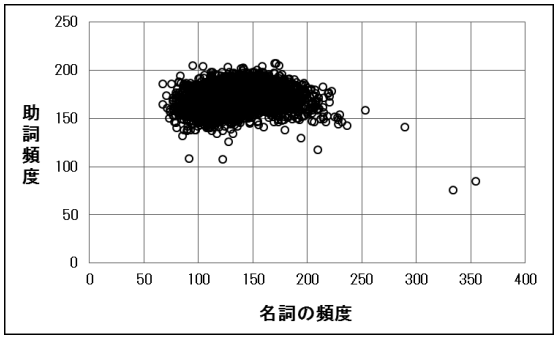

図 7 名詞頻度と助詞頻度の散布図：
図書 SC 二重選抜データ，：N=7,189

図 6 は<figureBlock>と<list>のタグが含まれている文書を除いた選抜データ，図 7 はここからさらに国立国語研究所（2015）の章節構造データのみを残した二重選抜データである。図 3 のフルデータから図 4 の章節構造データ，図 6 の選抜データ，図 7 の二重選抜データと削除数を増やすと，外れ値と思われるサンプルがより多く除かれていくことが確認できる。しかし，図 7 の二重選抜データでもなお外れ値と思われるサンプルが若干残っている。

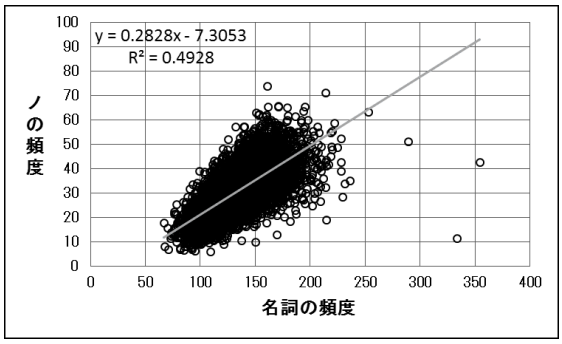

図 8　名詞頻度とノの頻度の散布図
図書 SC 二重選抜データ，：N=7,189

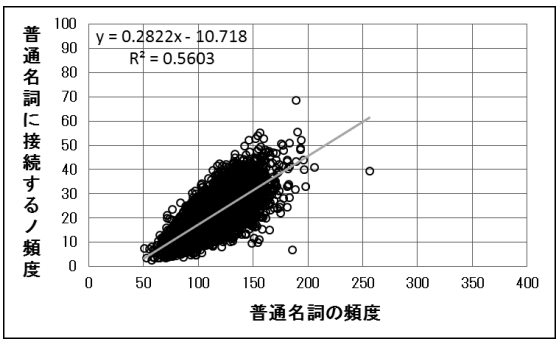

図 9　普通名詞頻度と普通名詞に接続するノの頻度の散布図，図書 SC 二重選抜データ，：N=7,189

図 8 は名詞とノの散布図で，二重選抜データを使用しているものの，このままでは，依然として外れ値の影響を受けることが考えられる。図 8 で名詞頻度が多いサンプルを観察すると，なお，固有名詞や数詞の列挙が残っていることが分かる。図 8 で最も名詞が多いサンプルは，先に挙げた例文（9）の『首都圏白書』で，2 番目に多いものが（12）の『立川飛行場物語』である。これらは，道路の路線名や町名などの固有名詞が列挙されているため，名詞数が多くなっている。

（12）大正十年の東京府電話帳を見ると、立川で五十三本の電話がひかれていたことがわかりますが、町名と氏名は次のようになっています。　＞零番＝立川郵便局―公衆通話用及び電報託送用　＞１番＝立川郵便局―一般事務用　＞　２番＝岩崎輝彌―子安農園立川分園―上古新田　＞　３番＝野沢源次郎―貿易商―下和田　＞　４番＝馬場福太郎―旅館―停車場前　＞　５番＝園部五郎吉―糸繭商―停車場前　＞　６番＝内藤九一―米穀商―停車場前　＞　７番＝旗野留五郎―雑貨商―停車場前　＞

８番＝村野安五郎―肥料商―停車場前　＞９番＝和知平三郎―雑貨商―停車場前＞十番＝中村久之助―料亭―停車場前　＞（BCCWJ サンプル ID：LBb3_00039，三田鶴吉，『立川飛行場物語』，500 語当たりの名詞頻度：333.8 語，500 語当たりのノ頻度：11.3 語）

そこで図 9 のように名詞を普通名詞に絞り，ノも普通名詞に接続するものに絞ると，概ね外れ値に影響されない状態になる。よって，分析に使用するデータは基本的に二重選抜データにし，調査対象は普通名詞と普通名詞に接続するノとすることにする。二重選抜データにおける普通名詞の頻度は 867,737 語で，全名詞の 79.1%，普通名詞とそれに接続するノの頻度は 162,769 語で，全ノの 70.5%，全名詞に接続するノの 80.3%になる。

３　分析結果と考察

３．１　外れ値と回帰直線の関係

本来はフルデータを使用し，全名詞と全ノの回帰直線を観察するのが望ましい調査である。しかし，本研究ではフルデータを 7 割弱に絞り込み，調査対象も普通名詞とそれに接続するノに限定することにした。本節ではなぜこのような絞り込みを行う必要があるのかについて，改めて説明する。

図 10 は，専門性や難易度でノの使用が変化するかどうかを調査するため，書籍の流通管理のために付与されている日本図書コード（C コード）の「教養・専門」と「児童」の区分を使用し，フルデータで普通名詞とそれに接続するノの散布図と回帰直線を描いた図，図 11 は，選抜データで同様の内容を描いた図である。

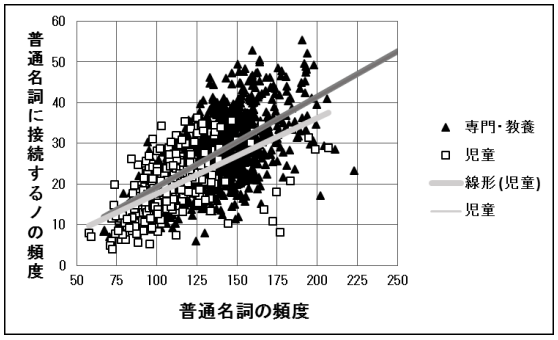

図 10 普通名詞とノの散布図・フルデータ

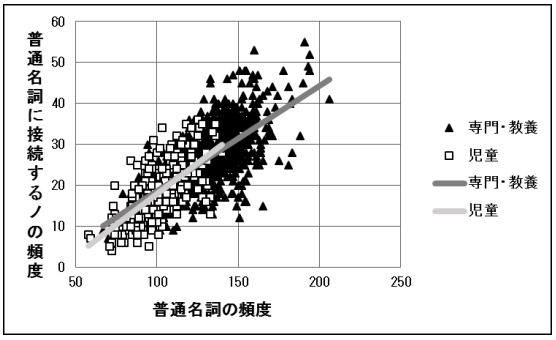

図 11 普通名詞とノの散布図・選抜データ

フルデータを使用した図 10 では，「児童」より「専門・教養」の傾きが急で，専門性の高い文体では，「児童」よりノが使用される割合が高いと判断される。しかし，選抜データを使用した図 11 では，回帰直線は一致し，専門性の高さによってノが使用される割合は変わらないと判断される。つまり，分析目的に不向きな文書を除けば，難易度や専門性が高いからと言ってノが多用されるわけではなく，専門性の高い書籍でも，児童向けの書籍でもノの使用傾向は一定であると考えられる。

選抜データに絞り込むために分析から除外した文書には，先に用例を示した（7）『昆虫』，（8）『世界遺産ガイド』が含まれており，（7）は「児童」，（8）は「教養」に分類されている。表 3 は，「児童」のフルデータ 387 文書から，60 の文書を除いた中で，普通名詞の数が多い文書 top10 のリストである。この第 1 位が用例（7）の『昆虫』である。これ以外の文

書も書名を見ると，図鑑，辞典，スポーツの解説書など，章節構造を持ったテキストとは明らかに異なる構造を持ったテキストであることが分かる。

表 3：「児童」から除いた普通名詞の多い文書 top10

ID	書名	普通名詞	ノ
LBqn_00015	昆虫	207	29
LBmn_00029	蛾蝶記	199	29
LBln_00025	道ばたの食べられる山野草	195	31
LBkn_00001	見てわかるルアーフィッシング	184	33
LBhn_00007	植物記	183	21
LBgn_00032	漢字事典五年生	177	8
LBpn_00009	服部幸應のはて・なぜ・どうしてたべものクイズ	174	18
LBnn_00001	バスケットボール	172	11
LBgn_00015	漢字事典四年生	167	14
LBdn_00020	New野球テクニック	166	35

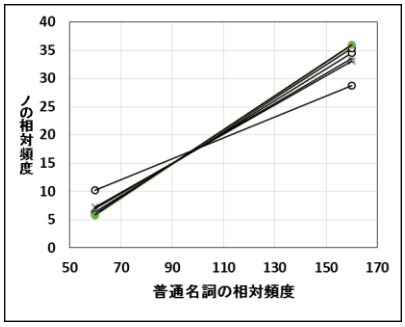

図 12：文書削除数別回帰直線

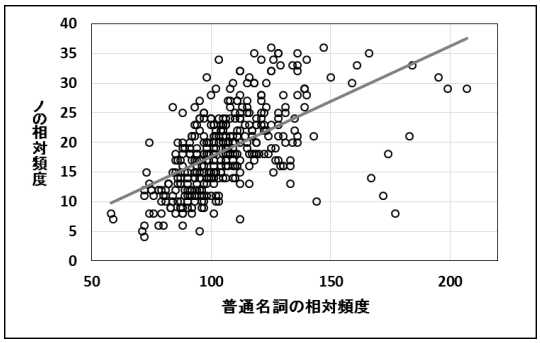

図 13 「児童」の散布図・フルデータ

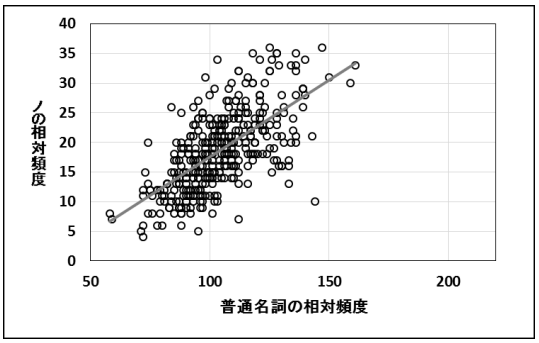

図 14 「児童」の散布図・10 文書削除

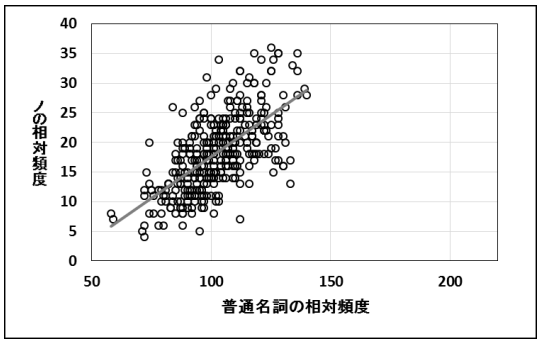

図 15 「児童」の散布図・30 文書削除

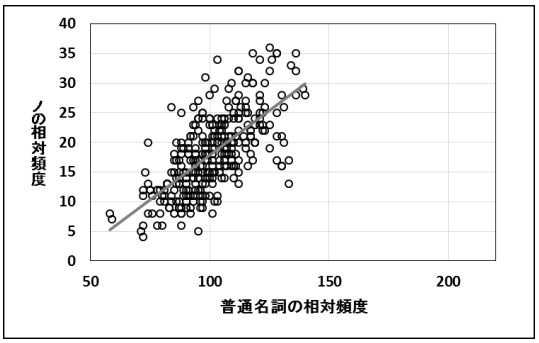

図 16 「児童」の散布図・選抜データ

ただし，<figureBlock>と<list>のタグが含まれている 60 文書をすべて除く必要があったかどうかの判断は難しい。図 12 はフルデータから普通名詞が多い順に特殊文書を 10 文書ずつ除いていった回帰直線 7 本の比較である。最も傾きが低いのがフルデータ，その上の直線がそこから表 3 の 10 文書を除いた時の回帰直線である。変化が激しいのは初めの 10 文書を除いた場合だけで，60 文書まで除く必要はなかったという考え方もできるかもしれない。

図 13 はフルデータで，明らかに外れ値と思われる文書が図の右側に散らばっている。図 14 はこれから 10 文書を除いた散布図，図 15 は 30 文書除いた散布図，図 16 は 60 除いた散布図で，これが選抜データとなる。散布図で確認しても図 14～図 16 の違いはごくわずかで

ある。しかし，文書を削除する基準をどこで線引きするかは難しく，恣意的なデータ操作を避けるためには分かりやすい基準に従うのが妥当だと思われる。

３．２　他の文体指標の結果と考察

「難易度や専門性」の違いによる普通名詞とノの回帰分析に続き，他の文体指標を使った調査の結果を示す。文体指標は，国立国語研究所（2015）を利用し，「硬度」「くだけ度」「客観度」「語りかけ性度」及び，章節構造データには分類されていない話し言葉の「対話系」と，それ以外の文書で，普通名詞とノの回帰直線がどのように異なるかを観察した。これらの指標は 2 段階～5 段階に区分されているが，図 17～図 21 では対極に位置する指標のみを使用している。分析データは図 19 のみ選抜データ，それ以外は二重選抜データを使用した。

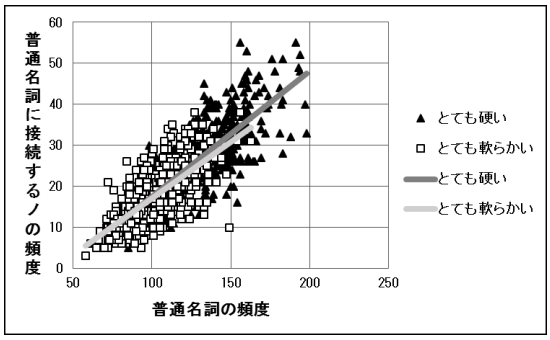

図 17 普通名詞とノの散布図・硬度

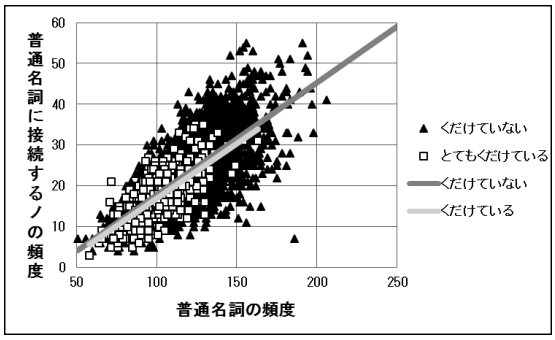

図 18 普通名詞とノの散布図・くだけ度

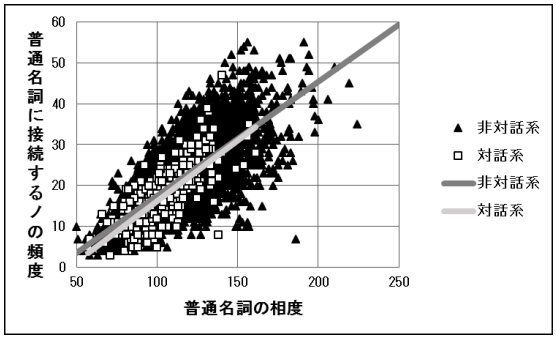

図 19 普通名詞とノの散布図・対話・非対話

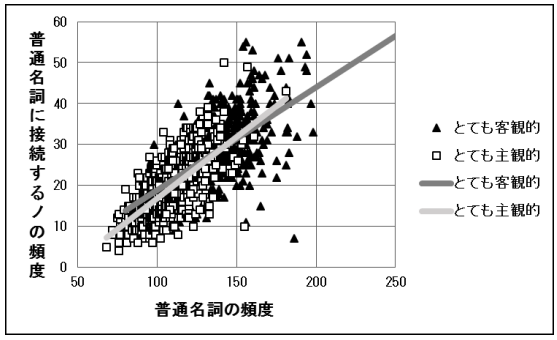

図 20 普通名詞とノの散布図・客観度

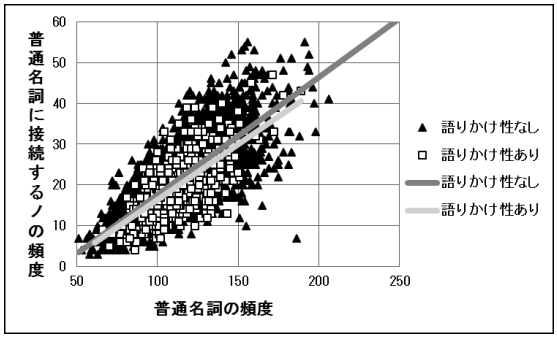

図 21 普通名詞とノの散布図・語りかけ性

表 4：文体別回帰式の係数と R^2

	傾き	切片	R^2
専門・教養	0.257	−7.122	.474
児童	0.301	−12.239	.451
とても硬い	0.307	−13.278	.530
とても軟らかい	0.274	−10.296	.475
くだけていない	0.276	−9.810	.520
とてもくだけている	0.268	−9.883	.465
非対話系	0.278	−10.304	.552
対話系	0.295	−13.197	.549
客観的	0.251	−6.249	.386
主観的	0.302	−13.337	.454
語りかけ性なし	0.287	−11.066	.584
語りかけ性あり	0.272	−10.809	.479

注：t 検定の結果すべての係数は 5%水準で有意

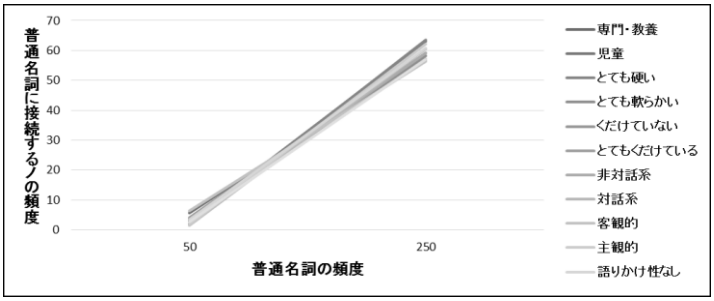

図 22　文体別回帰直線比較

図 17～図 21，及び表 2 を見ると，これらの指標の回帰式はほぼ一致し，文体指標や話し言葉・書き言葉（「対話・非対話」）で普通名詞とノの関連性は変化しないことが示唆された。文体によって普通名詞の頻度が特徴的な分布の違いを見せる一方，どのような文体であっても普通名詞の頻度が同じなら，使用されるノの頻度はほぼ同じになる。つまりノによって連体修飾節を作る述べ方は，難易度や専門性だけでなく，話し言葉や書き言葉，文章の硬軟，くだけ度，客観性，語りかけ性などには影響されない。これらの文体の違いは，執筆者の個性や執筆の意図の違いによって生じていると考えるのが自然である。このため，どのような文体であっても，普通名詞の頻度が決まればほぼ機械的にノの頻度が決まるという現象は，ノの使用に人間の個性や意志が介在する余地が小さいことを示唆していると考えられる。

４．まとめと今後の課題

本研究では BCCWJ 図書 SC のサンプルに文体指標をつけた国立国語研究所（2015）を利用し，「専門性」「硬度」「くだけ度」「客観度」「語りかけ性度」などの文体の違いによって，連体助詞ノの使われ方が異なるかどうかを調査した。

これを調査するには，テキストごとの名詞の頻度を X 軸（説明変数），ノの頻度を Y 軸（目的変数）とする散布図に回帰直線を描き入れて，この傾きや切片が文体の違いによって異なるのかどうかを観察すればよい。回帰直線の傾きや切片が同じであれば文体による使い分けはないと考えられる。

ただし，回帰直線は外れ値の影響を強く受けるため，恣意的ではない基準で，できるだけ外れ値を減らす方法を検討した。図書 SC のサンプルを観察すると，図表が含まれている文書や固有名詞・数詞が多用されている文書で，名詞の列挙が頻出する例が見られた。このため，文書構造タグの，<figureBlock>と<list>のタグがついている文書を除き，普通名詞と普通名詞に接続するノの頻度に絞って分析することで，外れ値の影響を受けにくい分析が行えると考えた。

<figureBlock>と<list>のタグがついている文書を除いた選抜データを使用し，国立国語研究所（2015）の文体指標を用いて分析すると，さらに対象となる文書が絞り込まれる。本研究ではこれを二重選抜データと呼ぶ。この二重選抜データを使用して，普通名詞と普通名詞に接続するノの頻度の回帰直線を描くと，文体の違いによって回帰直線の傾きや切片が異なることはなかった。

文体の違いは，執筆者の個性や執筆の意図の違いによって生じていると考えられる。このため，どのような文体であっても，名詞の頻度が決まればある程度機械的にノの頻度が決まるという現象は，ノの使用に人間の個性や意志が介在する余地が小さいことを示唆していると考えられる。

本研究ではコーパスの全データを使用せず，できるだけ外れ値が含まれないような基準を模索した。回帰分析において外れ値を除くことは重要だが，どのような方法を取れば，必要最小限のデータを除くことができるのか，今後さらに工夫していく必要がある。また，接続助詞やその他の助詞も文体に関係なく増減するのか，それを調査するためにはどのようなデータ選択を行う必要があるのか，これらを検討しながら調査を進めていくことが今後の課題である。

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，pp.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）.（http：//pj.ninjal.ac.jp/corpus_center/anno/よりダウンロード可能）

森秀明（2017）「一般的な日本語テキストにおける助詞比率の規則性」『言語資源活用ワークショップ 2017 発表論文集』，国立国語研究所，pp.9-22.

西部みちる・大島一・間淵洋子・小林正行・田島孝治・高田智和・山口昌也（2011）『『現代日本語書き言葉均衡コーパス』における電子化テキストの構築』，特定領域研究「日本語コーパス」平成 22 年度研究成果報告書（JC-D-10-03）.

山口昌也（2014）「第 3 章　文書構造の電子化」山崎誠（編）『講座日本語コーパス　2.書き言葉コーパス　設計と構築』朝倉書店，pp.45-67.

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる（2011）『『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2』，特定領域研究「日本語コーパス」平成 22 年度研究成果報告書（JC-D-10-04）.

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.