博士論文

コーパスの計量的分析法再考

東北大学大学院文学研究科言語科学専攻

森　秀明

目次

第 1 章　序論

本研究の目的は、コーパスを使用した計量的な言語分析において、これまで当然視されてきた基本概念や基本的な分析法を再考し、文字、単語、文などの言語単位を観察単位と考えてきたこれまでの分析法に替わって、統計学的にも言語学的にも有効な分析法を体系的に提案することにある。本章ではコーパスを使用した日本語学研究の現状（第 1 節）、本研究が必要とされている背景と研究の目的（第 2 節）、本研究の中心的主張（第 3 節）、本研究の構成（第 4 節）について述べる。

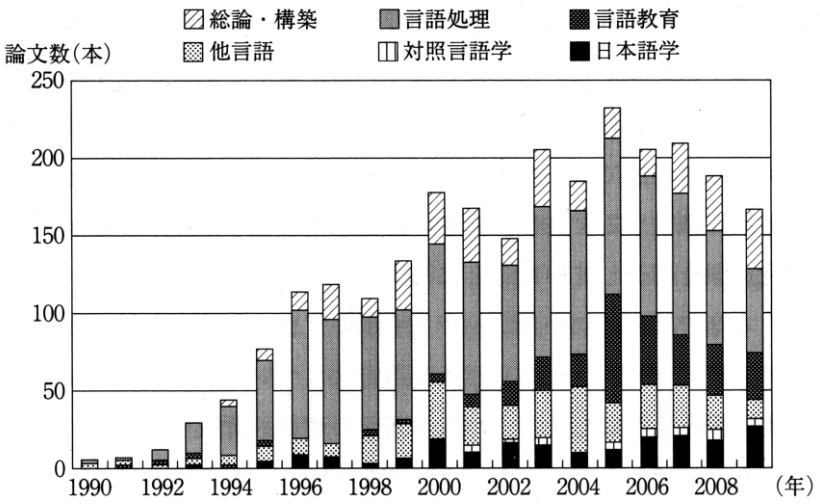
第 1 節　コーパスを使用した日本語学研究の現状

本節では、コーパスの定義を確認し、近年、コーパスを使用した言語研究が盛んに行われるようになってきた一方で、その分析方法については、度々問題点が指摘されている現状を概観する。

コーパスとは、言語研究のために大規模に集積された電子的な言語データのことである。石川（2012:13）ではコーパスの成立要件として、「(1) 書き言葉や話し言葉などの現実の言語を、(2) 大規模に、(3) 基準に沿って網羅的・代表的に収集し、(4) コンピュータ上で処理できるデータとして保存し、(5) 言語研究に使用するもの」という 5 点をあげている。この中で (2) の大規模性、(3) の代表性については、かなり幅があるのが現状で、同じコーパスでもこの二つの条件をある程度満たしている均衡コーパスと、特定の教育機関に所属する語学学習者の産出データを集めた小規模な学習者コーパスなどでは大きな違いがある。日本におけるコーパス研究の初期段階に、コーパスという概念を紹介した後藤（1995）では、上記 5 要件をほぼ満たす言語データを「狭義のコーパス」、いずれかの要件が十分でないものを「広義のコーパス」と呼び分けている。現在、狭義のコーパスで公開されているものは、無作為抽出によってサンプルを抽出している均衡コーパスが主体であるため、本研究でコーパスの成立要件を問題にする場合は、上記 5 要件をほぼ満たす言語データを「均衡コーパス」、それ以外のコーパスを「広義コーパス」と呼び、単にコーパスと呼ぶときはこの両者を含めたコーパス全体を指す。

近年、コーパスを使用して行われた日本語や日本語教育に関する研究（以後、これらを総称してコーパス日本語学研究と呼ぶ）を目にする機会が増えてきた。図 1.1 は間淵（2011:167）より引用したグラフで、1990 年から 2009 年までにコーパスを使用して研

究された論文数の推移を表している。この中でコーパス日本語学研究は 2000 年代に入った頃から増加傾向が顕著になり、コーパスの構築や言語処理の研究と並んで、毎年一定数の論文が発表されるようになってきた。



間淵（2011:167）図 1 より引用

図 1.1 コーパスを使用した論文数の推移

図 1.2 は 2010 年から 2016 年までの論文について、筆者が簡易的に調査・作成したグラフである。

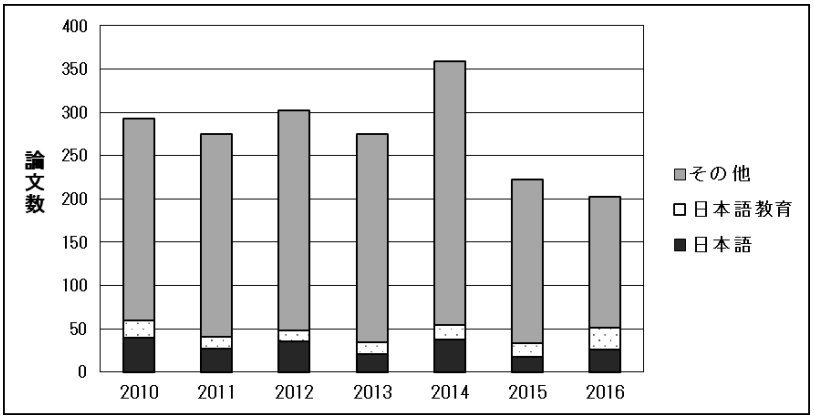


図 1.2 コーパスを使用した 2010 年以降の論文数の推移

図 1.2 は国立情報学研究所が Web 上で提供する「NII 学術情報ナビゲータ CiNii」を使用し、「コーパス」を検索語としてヒットした数（= a）をベースに、国立国語研究所が Web 上で提供する「日本語研究・日本語教育文献データベース」を使用し、分野を

「日本語教育」に絞り込んだ上で、「コーパス」を検索語としてヒットした数（= b）と、分野を「日本語情報処理」と「日本語教育」以外にして「コーパス」を検索語としてヒットした数（= c）を使用して描いた。図 1.2 の「日本語」の数は c、「日本語教育」の数は b、「その他」の数は a－b－c を表している。ただし、この二つの検索サイトの論文は完全な包含関係にはないため、厳密な調査にはなっていない。図 1.2 を見ると 2010 年以降もコーパスを使用した研究の総数は毎年 200 本以上にのぼり、コーパス日本語学研究の分野でも毎年 50 本程度の論文が発表されるなど、コーパスを使用した研究が根付いている現状が伺える。

しかし、その研究の方法については、いくつかの問題点が指摘されている。コーパス日本語学研究の黎明期から電子的な言語データの使用に関する問題点を指摘し、その後も継続してコーパス言語学の啓蒙と注意喚起を行った研究に後藤（1993，1995，1997，2003，2007 など）がある。これらの研究成果は多岐に渡るが、一貫して主張されてきた観点は、電子的な言語データを使用して一般化できる研究を行うのであれば、そのデータは研究対象とする言語に対し、代表性を持ったコーパスである必要があるということである。

世界ではじめて作られた均衡コーパスは 1964 年に完成した Brown コーパス（Brown University Standard Corpus of Present-day American English、概要は第 2 章第 2.1 項参照）である。一方、日本では 2011 年に公開された現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese : BCCWJ、概要は第 2 章第 2.2 項参照）が製作されるまで、代表性を持ったコーパスは存在しなかった。この間、さまざまな偏りのある電子的な言語データを使用した研究が行われたが、後藤の一連の研究はその時々に、コーパスを使用した日本語学研究の進むべき方向性を示してきた。ただし、BCCWJ が作られたからといって、すべての問題が解決するわけではない。後藤（2007:53）は BCCWJ の製作が始まることを記念して特集された論文集の中で、次のような注意喚起を行っている。

> このようにして近い将来に日本語のコーパスが広く使われるようになることは極めて望ましいことである。それを十分に活用するためには，それが存在するだけでは不十分であり，利用者の側にその活用に必要な知識と技能を得ようとする主体的な努力が要求される。コーパスは手軽に情報を得ることができるブラックボックスではないのであり，その性質を十分に理解した上で扱わな

ければ意味のある結論には結びつかないからである。　　（後藤，2007:53）

このようなコーパス研究に対する注意喚起は、伊藤（2005）でもなされている。伊藤（2005:89）では、「伝統的な計量言語学の成果を知らずに，新しいコーパス言語学に走る文系の研究者が多くなったため，現在「研究の質の劣化」という深刻な事態が進行しつつある」として、(1) 自分の研究と公開コーパスとのミスマッチ、(2) 形態素解析の精度の問題、(3) 自分の研究と市販の分析プログラムとのミスマッチの三つの観点から注意を促し、論文の最後を次のような言葉で結んでいる。

> コーパスを統計処理するときに，一番さけたいことは，コーパスの内容も知らず，プログラムの処理内容も知らないままで，それらしい統計データを出すことである．いわば，ブラックボックスのコーパスをブラックボックスのプログラムで処理するわけであるが，その場合，それを行っている人間はいったい何をしたことになるのか．それを調査や研究と呼べるのか．言語研究者が，読んだこともないテクストを研究することほど，矛盾に満ちたものはない．この状態が一般化することは，いわば言語研究が危機に陥っていくことを意味する．本稿が，そのような風潮に少しでも歯止めをかけることができれば幸いである．
>
> （伊藤，2005:96）

伊藤（2005）の指摘は自らデータを集め、自作のプログラムを組んで分析してきた研究者ならではの指摘であり、この水準を一般的な研究者の全てに求めることは難しいかも知れないが、コーパスの設計デザインを把握して自分の研究に使用することやコーパスに集積されているテキストの中身を確認して研究を行うことなどは、誰しもが行うべき分析法であるのは、確かである。

筆者は森（2017）において、日本語教育研究で最も多用されてきた KY コーパス（概要は第 5 章第 1.1 項参照）を使用した計量的な研究の概観を行った。そこでは、学習者ごとに語数の異なる言語データを使用しているにも関わらず、語数の平準化を行わないまま単語頻度の比較を行うなど、初歩的な統計のレベルで問題を抱えている研究が散見された。

伊藤（2005）や後藤（2007）などの注意喚起が度々なされているにも関わらず、コーパスの設計にそぐわない研究や統計学的に問題のある研究などが、いまだに行われているのがコーパスを使用した日本語学研究の現状といえるであろう。

第２節　研究の背景と目的

コーパス日本語学研究に関して分析方法の問題点が度々指摘される背景には、これまでコーパスを使用した計量的研究において、具体的にどのような方法を使用すれば有効な分析ができるのかについて、十分な議論が行われてこなかった点があげられる。そこで本研究では、コーパスを使用した言語分析においてこれまで当然視されてきた基本概念や基本的な分析法を再考し、文字、単語、文などの言語単位を観察単位と考えてきたこれまでの分析法に替わって、統計学的にも言語学的にも有効な分析法を体系的に提案することを目的にする。

はじめに、コーパスを使用した計量的な分析法において、どのような点が明確になっていないのかを、後藤（2007:54-5）で紹介されている「喫緊」という単語の調査を例に考えてみよう。後藤（2007:54-5）では、「喫緊」が出現するジャンルや共起する単語には著しい偏りがあるのに、国語辞書にはそれらの特徴が記されていないという問題意識をもとに、自らが作成したテキストデータベースを使用して、「喫緊」がどのようなジャンルにどれぐらい出現するかの調査を行った。その結果、「喫緊」が出現した 153 例の内、96 例が白書で、小説などにはほとんど出現せず、その大半にあたる 137 例が「喫緊の課題」という結びつき（コロケーション）で現れたという。これらの情報は国語辞書には載っていないため、「喫緊」という単語を新たに学習し、違和感のない場面で使用するには有益な情報である。しかし、後藤（2007:54-5）では、この調査で使用しているデータには問題があるとして、以下のように述べている。

> ここで使ったデータは，狭義のコーパスではなく，筆者がたまたま収集することのできたテキストの集合である。筆者が市販のテキストを個人的に収集したものであり，事前に全体を設計したものではない。これにはいくつかの決定的な欠点がある。これらはそもそも無原則的に集められたものであり，さまざまな位相の間での違いを印象以上に述べることが難しい。ここで言えることがどの程度まで現代日本語に対して一般化できるかは明らかではない。用例の実数を挙げてはみたものの，その数字にどれほどの意味があるのか，疑わしい。〔･･･〕。
>
> コーパスが整備されることによって，ここで行ったような記述がより精緻化され，積み重ねられていけば，語彙項目間に見られる関連や文法現象との関連に対するより深い理解につながることが期待でき，さらには語義のより深い分

析や，文法や語用論の面のコーパス言語学も次第に整うであろう。

（後藤，2007:54-5）

現在は均衡コーパスの BCCWJ が完成しているため、この調査を追試することができる。表 1.1 は BCCWJ を使用して「喫緊」を検索した結果である。

表 1.1　BCCWJ・短単位を使用した「喫緊」の頻度比較

サブコーパス	レジスター	固定長頻度	統合形式頻度	固定長調整頻度	統合形式調整頻度	固定長語数	統合形式語数	サンプル数
図書館SC	図書館書籍	1	3	0.15	0.10	6,702,069	30,377,866	10,551
出版SC	出版書籍	3	14	0.47	0.49	6,387,438	28,552,283	10,117
	雑誌	0	1	0.00	0.22	1,162,449	4,444,492	1,996
	新聞	3	3	3.22	2.19	930,928	1,370,233	1,473
特定目的SC	白書	3	18	2.88	3.69	1,041,914	4,882,812	1,500
	Yahoo！知恵袋		0		0.00		10,256,877	91,445
	Yahoo！ブログ		3		0.29		10,194,143	52,680
	国会会議録		25		4.90		5,102,469	159
	広報誌		5		1.33		3,755,161	354
	ベストセラー		0		0.00		3,742,261	1,390
	法律		0		0.00		1,079,146	346
	教科書		0		0.00		928,448	412
	韻文		0		0.00		225,273	252
合計		10	72	0.62	0.69	16,224,798	104,911,464	172,675

BCCWJ の設計については第 2 章で詳述するが、BCCWJ には 3 種類のサブコーパス（以下 SC と略す）がある。このうち図書館 SC と出版 SC が母集団を定めてデータを無作為抽出した SC、その他に多様なレジスター（媒体）を集積した特定目的 SC がある。集積したデータの長さには 2 種類あり、文字数を約 1,000 字に固定して集積したデータが固定長、章や節などのまとまりに合わせ、長さを変えて集積したデータが可変長である。本研究では固定長と可変長を統合し、重複を除いたデータを統合形式と呼ぶ[1]。また、形態素解析を行う言語単位には、意味を持つ最小の単位をもとに規定した「短単位」と、文節をもとに合成語や複合辞を 1 単位に規定した「長単位」があり、「短単位はコーパスからの用例収集に適した単位であり、長単位は BCCWJ に格納したレジスターの言語的特徴の解明に適した単位である」とされている（国立国語研究所コーパス開

[1] BCCWJ のマニュアルである国立国語研究所コーパス開発センター（2015）では、固定長や可変長の説明箇所に「統合形式」という名称は使用されていないが、p.160 には、「形態論情報付き統合形式 XML（Morphology-base XML 以下、M-XML と略記する）は、文字ベースの XML（C-XML）フォーマットをもとにして、固定長・可変長サンプルを統合し、言語構造を一定程度反映させた XML フォーマットである」とある。また、『現代日本語書き言葉均衡コーパス』語彙表 ver.1.1 解説には、「統合形式とは、重複のないように固定長と可変をあわせたものである」との注記があるため、「統合形式」という名称を使用する。

発センター，2015:26)。これらの数はそれぞれ短単位数、長単位数と呼ぶのが正確であるが、本研究では簡略化して「語数」と呼ぶ。表 1.1 は、短単位を使用して集計している。

BCCWJ は総計 1 億語のコーパスである。しかし、均衡コーパスと呼べるのは図書館 SC と出版 SC の固定長だけだといわれている（田野村，2014:121-3)。図書館 SC と出版 SC の統合形式は、個々のテキストの長さがばらばらであるため、均衡コーパスとは呼びにくい。これ以外の特定目的 SC は日本語を代表するのに不可欠なレジスターであるから選ばれたというより、どちらかといえば後藤（2007:55）がいうところの「無原則的に集められた」データに近い。特に分量が多い Yahoo！知恵袋、Yahoo！ブログ、国会会議録などのデータ量は、その分量が日本語を代表するのに適量だから集積されたというより、元々のデータが電子化されていたため、コストをかけずに集積できるという観点から分量が多くなったと思われる。

田野村（2014:121-3）に従うなら、表 1.1 の図書館 SC と出版 SC の固定長が母集団に対して代表性を持つ頻度である。これらの頻度はごく低いが、果たしてこの頻度を信頼してもよいのであろうか。それともこれほどの低頻度の場合、まだしも語数が多い統合形式の頻度の方が正確なのだろうか。また、これらの頻度を比較する場合、そのままの頻度を比べてもよいのだろうか。それとも何らかの調整を施す必要があるのだろうか。

BCCWJ はその構築に当たって、詳細な報告書が 11 冊作成されている（丸山・秋元，2007；丸山・秋元，2008；柏野・丸山・稲益・田中ほか，2009；丸山・山崎・柏野・佐野ほか，2011a；丸山・山崎・柏野・佐野ほか，2011b；高田・小林・間淵・大島ほか，2009；西部・大島・間淵・小林ほか，2011；山口・高田・北村・間淵ほか，2011；小椋・小磯・冨士池・宮内ほか，2011；小木曽・中村，2011)。また、マニュアルに当たる『『現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版』（国立国語研究所コーパス開発センター，2015）や、これらの報告書や利用の手引きの内容をコンパクトにまとめた解説書である山崎誠（編）（2014）が存在する。しかし、それらのどこを読んでも、表 1.1 の固定長や統合形式の頻度をどのように調整したり解釈したりすればよいのかについての実際的な説明は書かれていない。

固定長は統計的な分析に向き、可変長はテキストの論理構造の把握や文体の調査などに向くという記述はあるが（丸山・柏野，2014:26；国立国語研究所コーパス開発センター，2015:30)、可変長を使用して計量的な分析を行ってよいかどうかの記述はない。また、固定長と可変長の大きな違いはテキストの文字数を一定にしているか、大きな幅

を持たせているかという点にあるが、固定長にしてみても文字数を約 1,000 字に固定しただけであり、これを語数に直すと個々のテキストの長さは媒体によってかなり異なる。

表 1.2 は、固定長が備わっている五つの媒体の平均語数を比較した表である。最も語数が少ない媒体に比べ最も語数が多い媒体の語数は、短単位でも長単位でもどちらも 1.19 倍になっている。2 割弱ほどサイズが異なるデータ同士をそのまま比較するのは問題があるようにも思われるが、先に挙げた報告書類には、それに対してどのように対処すればよいかの記述はない。

表 1.2 BCCWJ 固定長の平均語数

	短単位平均	長単位平均
図書館書籍	635.2	523.6
出版書籍	631.4	504.2
雑誌	582.4	458.8
新聞	632.0	455.1
白書	694.6	440.1

それでは、コーパス研究の基礎的な知見をまとめたコーパス言語学の概説書や、言語研究のための統計の概説書などを参考にすればどうであろうか。これらを読むと、サイズが異なるコーパスの頻度を比較する際は、調整頻度を算出すればよいと書いてある（バイバー・コンラッド・レッペン，2003:38-41；石川・前田・山崎（編），2010:27-8；石川，2012:114-5；マケナリー・ハーディー，2014:74-6 など）。調整頻度とは調査対象の頻度をコーパスの総語数で割って使用率を求め、これに一定数をかけて扱いやすくした頻度である。表 1.1 では使用率に 100 万語をかけ算し、100 万語当たりの調整頻度を計算している。ただしこれらの概説書には、固定長と統合形式のような二種類のデータが存在する場合、どちらが統計分析に適するのかについては書かれていない。

後藤（2007:54-5）の調査では、新聞などより白書の方が「喫緊」の頻度が高かった。表 1.1 では統合形式調整頻度で比較するとこれと同じ結果になるが（白書 : 3.69，新聞 : 2.19）、固定長調整頻度で比較すれば、反対の結果になる（白書：2.88，新聞：3.22）。このどちらの結果を信頼すればよいのかについて、明確な考え方を示した研究は管見のかぎり存在しない。均衡コーパスを使用する目的は日本語に対して一般化できる調査を行うことにあったはずだが、これほど基本的なことさえよく分からないのがコーパス言語学の現状である（この問題については、第 4 章で検討する）。

以上の例で分かることは、コーパスを新たに使い出した研究者の増加によって、研究

の質の低下が起きているとは言い切れない実情が存在しているということである。すなわち、研究の質を確保するための基本的な方法が、これまで十分に議論されてこなかったところに、真の原因が存在していると考えられる。

そこで本研究では、コーパスを使用した計量的な言語分析において、これまで当然視されてきた基本概念や基本的な分析法を再考し、文字、単語、文などの言語単位を観察単位と考えてきたこれまでの分析法に替わって、統計学的にも言語学的にも有効な分析法を体系的に提案することを目的とする。

本研究で分析に使用するコーパスは、日本語の研究で最も使用される機会が多いと思われる BCCWJ と、これまでの日本語教育研究で最も多く使用されてきた KY コーパス、および、今後の日本語教育研究で最も多く使用されると考えられる多言語母語の日本語学習者横断コーパス（International Corpus of Japanese as a Second Language：I-JAS）である。本研究で考察する内容は、これらのコーパスに限定されるものではないが、コーパス日本語学研究で多用されるコーパスを例に議論を行えば理解されやすく、今後これらのコーパスを使用した研究が行われる際にも、有益な情報提供ができると考える。

第 3 節　本研究の中心的主張

本研究の最も中心的な主張は、これまでの文字、単語、文などの言語単位を観察単位と考えてきた分析法に替わって、文書を観察単位とした言語分析を行えば、統計学的な意義や言語学的な意義が明確で、有効な分析が行えるという点にある。

統計学的に有効な分析を行うためには、次の 3 点に留意する必要がある。

①母集団から無作為抽出された母集団の構成要素が個体である。
②個体は独立していなければならない。
③統計分析の目的は、個体の観測値の分布からデータの特徴や性質をつかむことである。

この重要性を理解するために、統計分析の基本を述べた次の 3 つの引用を見てみよう。

> 統計的な調査の対象を一般に**母集団**と呼び, それを構成する各要素を**個体**と呼ぶ. 各個体に対して何らかの調査や測定が行われ, その特性を表す観測値（測定値，データなどともいう）が得られる．〔･･･〕観測値は個体ごとに変化するのがふつうであり, そのような観測値をひとまとめにして**変数**または**変量**とい

う〔･･･〕.

標本調査で得られた観測値から母集団のさまざまな統計的性質を合理的に推測することが数理統計学の目的である.

（尾畑，2014:1-2　注：太字は原文ママ、以下同じ。）

ここでは、標本調査で得られた個体の観測値から母集団の性質を推測することが統計分析の目的であると記されている。重要なのは「観測値は個体ごとに変化する」という点である。これを別の言い方で述べれば「分布する」という。次の引用は分布に関する引用である。

（筆者注：図表 1-1 女子大生 80 人の身長（㎝）は）「日本人の成人女性」の一部という集団を扱っていますが、属するメンバーの身長は、さまざまな数値をとります。この「**さまざまな数値をとる**」ということを、専門の言葉で「**分布する**」といいます。分布が生じるのは、その数値が決まる背後に何らかの「**不確実性**」が働いているからに、ほかなりません。不確実性のメカニズムが、まちまちな身長の数値を生み出すと考えるのです。ところが、「不確実」と一口にいっても、それらには固有の「特徴」や「癖」があることがわかっています。その固有の特徴や癖を「**分布の特性**」と呼びます。〔･･･〕。

そこで、この生データ、つまり「生の現実」から、何かその分布の特徴や癖を引き出すための手法が必要になります。それが「**統計**」という手法なのです。

（小島，2006:17）

この引用では、個体の観測値が分布するからこそ、有効な統計分析が行えるという統計の原理が分かりやすく述べられている。個体の性質で重要なのは、この「分布する」ということと、次の引用で説明されている、「独立している」ということである。

同じ条件に属す個々の対象は，本来，独立変数に関して同じ操作が施されていること以外は，なんらかの偏った（一定の）影響を受けていたり，なんらかの共通の特徴をもっていたりしてはいけません。〔･･･〕このような問題のある状態に陥っていることを，“個々のデータ（観測値）が独立していない”などといいますが，この“データの独立性”という条件は，すべての統計的検定に共通した，常に留意すべき重要な前提条件です。　（吉田，2001:248）

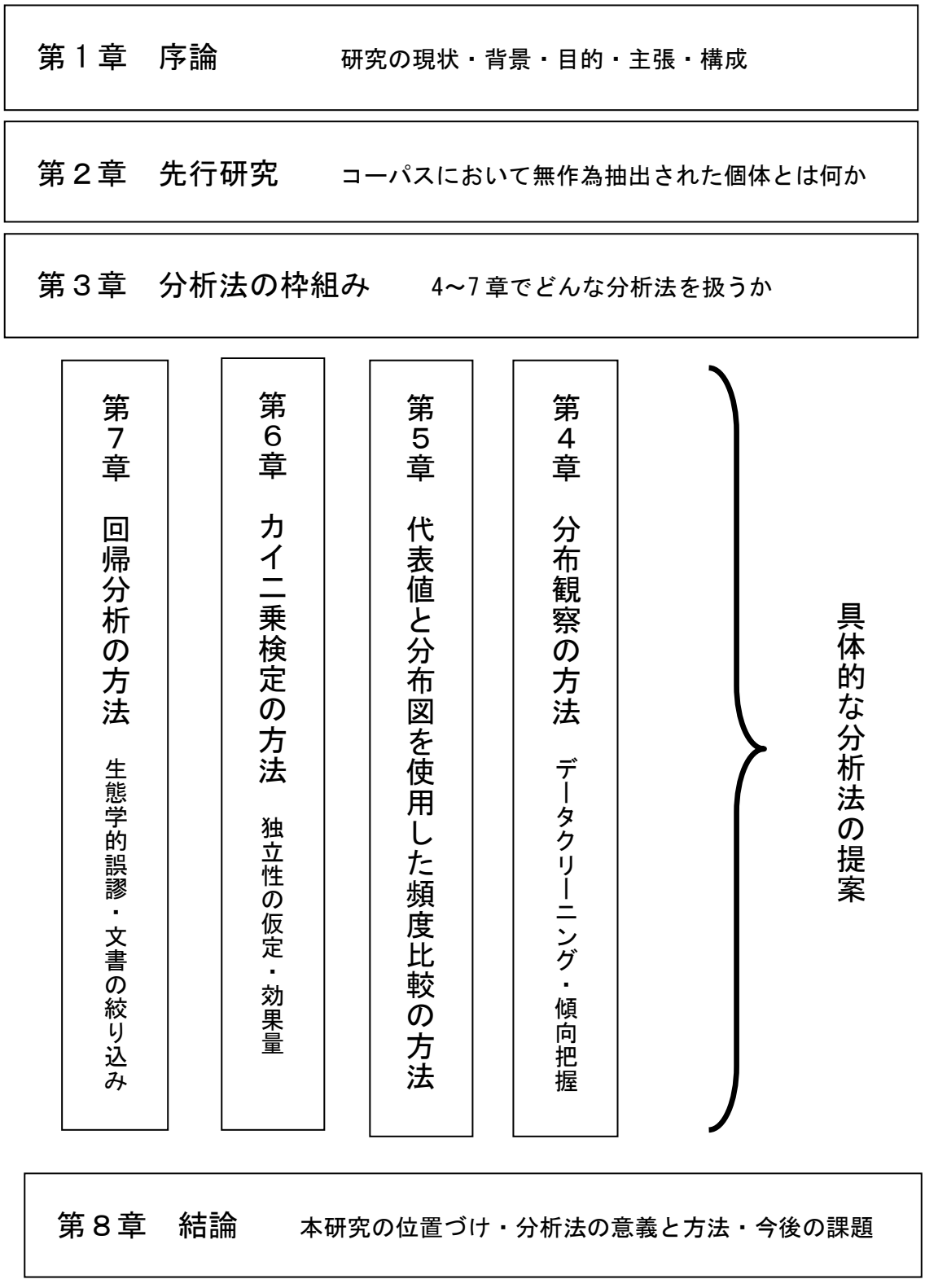
それでは、文字、単語、文などの言語単位を観察単位とした場合の観測値とは何だろうか。たとえば、コーパスの中にある「喫緊」という単語を観察した時、どの「喫緊」という個体も、それが出現したという点から考えると観測値は 1 である。単語と言う個体の観測値は分布しない。しかし文書なら、文書 A には「喫緊」が 1 回、文書 B には 3 回出現したなどのように、観測値はさまざまに分布する。観測値が分布するからこそ、母集団の性質が統計的に推測できるのであり、分布しない観測値で合理的な統計分析を行うのは困難である。

また、文字、単語、文などの言語単位は独立していない。独立とは任意の i 番目と j 番目のデータに関して「j 番目の分布が i 番目の値に影響されない」ということである（豊田（編著），2009:26）。人間は、一定の法則に従って、言葉を話したり、文章を書いたりしている。日本語であれば、名詞の後には助詞が出現しやすいという文法の制約もあれば、「喫緊の」の後には「課題」が出現しやすいというコロケーションの制約もある。これは、単語が独立していないことの証拠である。一方、文書であれば、コーパスの標本の一つとして文書 A が選ばれたからと言って、文書 B の選択には何の影響も与えない。コーパスで独立しているのは文書である。

観察単位とは、標本を作る際に無作為抽出した個体の単位のことである。文書を観察単位として分析を行うということは、無作為抽出された個体を文書と考え、文書から得られた観測値を使用して統計分析を行うことを意味する。文書を観察単位として統計分析を行うと、研究目的にそぐわない文書を排除することが容易で、言語研究の目的に適合した析ができる。また、特徴的な観測値を示す文書の中身を確認しながら分析できるため、分析結果の解釈も行いやすい。「文書を観察単位とした言語分析を行えば、統計学的な意義や言語学的な意義が明確で、有効な分析が行える」ということが、本研究の最も中心的な主張である。

第４節　本研究の構成

本研究の構成を簡単に図示すると以下のようになる。



本章の「序論」に続き、第 2 章の「先行研究」では、コーパスにおいて無作為抽出された個体は文書と考えられるため、言語単位で分析を行ってきたこれまでの研究方法には問題があり、文書単位の分析法を確立していく必要があることを述べる。

第 3 章の「文書を観察単位とした分析法の枠組み」では、本研究で扱う分析法について検討する。統計分析の解説書では、度数分布の観察→平均値や中央値などの代表値を使用した分析→t 検定やカイ二乗検定などの統計的検定→因果関係の解明などを目指した探索的な分析、の順番で記述されているのが一般的である。このため、本研究においても基本的にこの流れに従って検討していくこととする。

第 4 章の「分布観察の方法」では、文書度数分布図や散布図を描くことによって、特異な特徴を持つ文書や調査対象の全体的な分布傾向を観察する分析法について述べる。これにより、データクリーニングが容易になり、調査対象の特徴もつかみやすくなる。また、文書内の単語分布を観察することにより、文書を観察単位と考えた場合、固定長と統合形式のどちらが正確なデータであるといえるのかについても考察する。

第 5 章の「代表値と分布図を併用した頻度比較の方法」では、学習者コーパスを対象として、調整頻度、平均値、中央値などの代表値を使用した分析法の有効性を検討する。この結果、学習者コーパスはばらつきが大きいため、単独の代表値ではデータの特徴をうまく要約することが難しいことが明らかになる。そこで、代表値を使用した分析法に替えて、蜂群図という散布図に中央値と四分位点を描くことができる箱ひげ図を重ね書きして観察する分析法を提案する。

第 6 章では、コーパス言語学で最も多用されてきた「カイ二乗検定」を取り上げる。これまで行われてきたカイ二乗検定は、言語単位を観察単位にしているため、統計分析の前提となる独立性の仮定を満たすことができず、効果量を有効に評価することも難しかった。これに替わって文書を観察単位にした場合、独立性の仮定を満たすだけでなく、効果量も質問紙調査や実験などと同様の目安で評価できるため、統計学的にも言語学的にも有効な分析が行えることを述べる。

第 7 章は統計分析において因果関係の解明に最もよく使用されている「回帰分析」を取り上げる。言語単位を観察単位とした分析法では、媒体、ジャンル、学習者の習得レベルなどの集団を分析単位にして回帰分析を行うことが一般的であった。しかし、集団レベルの回帰分析では、個体単位の相関は低いのに、あたかも高い相関関係があるかのように誤認する生態学的誤謬を犯す危険性がある。また、集団レベルの分析の場合、本来なら分割して分析するべき分割相関に気づかないまま、誤った推論を行う可能性があ

る。第 7 章ではこのような生態学的誤謬や分割相関を見逃して分析した結果、誤謬が起きる例を示し、回帰分析においても個体単位で分析する重要性を述べる。また、文書を観察単位とした場合、分析目的にふさわしくない文書を除くことで、正確な回帰分析が行えることを示す。

第 8 章では、「結論」を述べる。本研究の目的は、コーパスを使用した言語分析においてこれまで当然視されてきた基本概念や基本的な分析法を再考し、文字、単語、文などの言語単位を観察単位と考えてきたこれまでの分析法に替わって、統計学的にも言語学的にも有効な分析法を体系的に提案することにある。本研究の問いは「コーパスを使用した計量的な言語分析において、どのようにすれば統計学的にも言語学的にも有効な分析ができるのか」ということであり、その答えを 3 点に要約して述べると、「①個体（文書や学習者）を観察単位として分析する、②分布図という地図を作って分析する、③分割相関や外れ値にかく乱されないで分析する」ということである。

本研究によって、これまで行われてきたコーパス研究の中には、必ずしも有効な分析になっていない研究が存在する可能性が示唆される。本研究では従来の分析法に替わる具体的な分析法の提案を行うため、この分析法を使用して、各研究者自らが過去に行った研究の再分析を行うことが可能である。本研究の意義と成果は、これまで当然視されてきた基本概念や分析法の中にも問題があることを明確にし、それに替わる具体的な分析法を提案する点にある。コーパスを使用した言語分析は、日本語学や日本語教育学において欠かせない研究分野に成長してきた。本研究は、このコーパス日本語学研究の分野に対し、統計学的にも言語学的にも有効な分析法を体系的に提案することで貢献を行う。

第 2 章　先行研究

本章では、これまでに行われてきた文字、単語、文などの言語単位を観察単位とした統計分析が必ずしも有効な分析法にはなっていないという問題点を明らかにし、これに替わる分析法が必要とされていることを述べる。

第 1 節では、コーパス構築にかかわる重要概念である代表性と無作為抽出ついて、先行研究における定義を概観する。第 2 節では、均衡コーパスとして構築された Brown コーパスと BCCWJ の設計方針を確認し、現在のコーパス言語学で文字、単語、文などの言語単位を観察単位にした分析が行われている根拠が、集落抽出法にあることを述べる。第 3 節では、言語データを集落抽出した場合、単語や文は独立していないため、有効な統計分析ができないことを述べる。第 4 節では、学習者コーパスを取り上げ、学習者コーパスにおける独立した個体は学習者であるため、学習者を観察単位とすべきことを述べる。第 5 節ではコーパス言語学に隣接する学問分野の研究方法を概観し、文体統計学や自然言語処理の分野では、文書を観察単位にした研究が早くから行われている現状を述べる。第 6 節では本章の結論として、先行研究の問題点と解決すべき課題をまとめる。

第 1 節　代表性と無作為抽出の定義

本節ではコーパス構築にかかわる重要概念である代表性と無作為抽出の定義について確認する。無原則に集められたデータとは異なり、均衡コーパスに信頼性があるのは、コーパスに代表性が備わるような設計がなされているからである。マケナリー・ハーディー（2014:361）では、代表性の定義について以下のように記されている。

> 代表性（representativeness）：各種のテキストタイプが現実の構成比と同等の比率で含まれるようにコーパスの標本が抽出されていること。そうしたコーパスは、代表性を持つコーパスと呼ばれる。コーパス内容が標本抽出の元となる言語ないし言語変種の全体を正確に反映する上で代表性は不可欠である。
>
> （マケナリー・ハーディー，2014:361）

次に、無作為抽出の定義を確認する。次の引用は、数理統計学の立場から記述された尾畑（2014）による「無作為抽出」の定義である。

有限母集団に対して，どの個体も等確率で取り出されるような標本抽出を**無作為抽出**という．このように取り出された標本を**無作為標本**，または，無作為は当然のこととして単に**標本**という．無作為標本に対して得られる観測値は，取り出された標本ごとに異なり，その値の現れ方は母集団分布に従う．つまり，無作為標本は母集団分布に従う確率変数とみなされる．　（尾畑，2014:160）

このような代表性と無作為抽出の関係について、山崎・前川（2014:5）では、以下のように記されている。

代表性はコーパスに求められる基本的かつ重要な性質である．「代表性を持つ／代表性がある」とは，コーパスが母集団の過不足のない縮図となっていて，コーパスから得られた観測値で母集団の状況を一定の精度で推測することができることを意味する．代表性を実現するための方法がランダムサンプリングである．　（山崎・前川，2014:5）

同様の記述は、前川（2013:13-5）にも見られる。

母集団が明確に決定できるのであれば　〔･･･〕　母集団を構成する全てのサンプルが等しい確率で選ばれる条件下でサンプルを無作為抽出することが代表性を保証する最も確実な手段である．　（前川，2013:14）

BCCWJ のマニュアルには以下の引用のように記されており、BCCWJ の製作に当たっては無作為抽出が重要視されたことが伺われる。

BCCWJ は日本語に関する初の均衡コーパスであるが、その設計にあたっては、先行する諸外国の均衡コーパスを参考にしており、いくつかの点で先行コーパスに優った設計がなされている。たとえば、厳密な無作為抽出を可能なかぎり実施していること（第 3 章参照）、平均サンプル長を British National Corpus などに比べると短めに抑えることによって文献による語彙の偏りを低減していることなどである。　（国立国語研究所コーパス開発センター，2015:1）

以上のように、代表性とはコーパスが母集団の過不足のない縮図となっていることで

あり、それを実現するために母集団から個体が等確率で選ばれる無作為抽出が重視されてきたことが分かる。問題は、何が無作為抽出されているのか、つまり、コーパスにおける個体とは何かである。

第 2 節　コーパス構築における無作為抽出の実際

本節では、コーパスを構築する際、何が実際に無作為抽出されているのかを中心に、第 2.1 項で Brown コーパスの設計を、第 2.2 項で BCCWJ の設計を確認する。Brown コーパスや BCCWJ は均衡コーパスと呼ばれ、コーパスに母集団の代表性を持たせるため、無作為抽出によって標本が抽出されているコーパスである。

第 2.1 項　Brown コーパスの設計と無作為抽出法

はじめに、Brown コーパスでは、何が無作為抽出されているのかを検討する。Brown コーパスは、1964 年に世界ではじめて構築されたコーパスである。米国ブラウン大学の W. N. Francis と Henry Kučera が、アメリカ教育省の研究資金を得て開発した。Brown コーパスでは標本の収集基準が詳細に定義されており、その後この基準を踏襲した Brown ファミリーと呼ばれる一連のコーパスが作られたこともあり、現在でも言語研究に幅広く使用されている。その設計の詳細は Brown コーパスのオンラインマニュアルに記されている（Francis & Kučera，1979）[2]。

Brown コーパスが代表性を持つといわれている最大の理由は、アメリカ現代英語を想定母集団とし、層化抽出法と無作為抽出法を組み合わせることで、均衡なコーパスを製作したことにある。想定母集団とは、「アメリカ現代英語」のように、その総体を捉えきれない母集団を意味する。これに対し、図書館の書籍リストのように範囲が確定できるものを現実母集団と呼ぶ（石川，2012:22-3）。

Brown コーパスでは、はじめに「現代アメリカ英語」の範囲を明確にするため、データを 1961 年に刊行された書き言葉の出版物に限ること、特殊な言語使用がなされている詩、話し言葉性が強い劇、会話が 50%を超える小説は対象外にする方針が立てられた。その後、カテゴリー別に現実母集団をブラウン大学図書館や Providence Athenaeum の蔵書目録と定め、無作為抽出法によって標本を決定した。しかし実際に使用されたのはこの二つだけではなく、新聞の目録にはニューヨーク公立図書館のマイクロフィルムが、

[2] http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM（2018.06.21 閲覧）。このマニュアルは、1964 年版をもとに 1971 年に改訂されたものを、1979 年にさらに改訂したものである。

特定雑誌の選択にはニューヨーク最大の古書店の蔵書が使用された。

カテゴリーはまず、情報散文と創作散文の 2 種類に大別され、その下位に 15 カテゴリーが置かれる。それぞれのカテゴリーには現実母集団の比率に応じた重み付けがなされている。サンプルはその重みに応じた比率で抽出され、それぞれほぼ 2,000 語のテキストを全 500 文書、計 100 万語のデータが集積された。表 2.1 は、カテゴリー別にテキスト数とその割合を記した表である。

表 2.1　Brown コーパスの構成と文書数

Francis & Kučera（1979）をもとに作成

情報散文				創作散文			
大ジャンル	カテゴリー	文書数	%	大ジャンル	カテゴリー	文書数	%
新聞	A:新聞雑誌・報道	44	8.8%	小説	K:一般小説	29	5.8%
	B:新聞雑誌・論説	27	5.4%		L:推理小説	24	4.8%
	C:新聞雑誌・評論	17	3.4%		M:SF小説	6	1.2%
一般散文	D:宗教	17	3.4%		N:冒険小説	29	5.8%
	E:技術・趣味	36	7.2%		P:恋愛小説	29	5.8%
	F:一般実用	48	9.6%		R:ユーモア小説	9	1.8%
	G:随筆・伝記・回顧録	75	15.0%				
	H:雑（政府文書など）	30	6.0%				
学術	J:教養系・科学系	80	16.0%				
	小計	374	74.8%		小計	126	25.2%
全合計				500（100%）			

Brown コーパスの特徴としては、母集団の割合に応じてカテゴリーの割合を定めていること、各サンプル数を 2,000 語に均一化していること、書き言葉を対象とするため、話し言葉性の強い劇や会話の比率が 50%を超える小説は対象外としていることなどがあげられる。

無作為抽出の詳細な方法は下記のとおりである。

> Once these categories, subcategories, and numbers of samples had been decided upon, the choice of the actual samples was made by various random methods, chiefly the use of a table of random numbers applied to the total list of available publications in the subject field in question. The page on which to begin the sample was also selected by the random number table. Each sample begins with the first complete sentence on the page so selected. Titles and running heads have been omitted, also footnotes, tables, and picture captions. A rough count of 2,000 words was made and the sample

was terminated at the next sentence-break.　（筆者訳：一度これらのカテゴリー、下位カテゴリー、およびサンプル数が決定されると、実際のサンプルの選択は、さまざまなランダムな方法によって行われた。問題となっている対象分野の利用可能な出版物の全リストに対しては、主に乱数表の使用を適用した。サンプリングを開始するページも、乱数表によって選択された。各サンプルは、このようにして選択されたページの最初の完全な文から始まる。タイトルと欄外見出し、さらに、脚注、表、画像の表題は削除される。大まかに 2,000 語の単語が選択され、サンプルは次の文末で終了された。）　（Francis & Kučera，1979）

大枠の設計基準に基づいて選別するところまでは、主観的な判断も交えて抽出基準が作られたが、この基準によって出版物のリストが選定されてから以降は、厳密な無作為抽出が行われた。まず、出版物のリストから乱数表を使用して任意の出版物を選び、さらにその出版物の任意のページを乱数表で選んで、そのページで完全な文として始まる先頭の文を起点として 2,000 語を超えた文の終結部までを抽出した。Brown コーパスの想定母集団は 1961 年に刊行された書き言葉の出版物の総体、現実母集団は図書館の蔵書目録などから選定した出版リストである。個体はそのリストを構成する書籍や新聞雑誌などの個々の出版物で、最終的に抽出されたのは個々の出版物から抜粋された約 2,000 語のテキストである。本研究ではこのようにコーパスの標本抽出基準に基づいて抽出されたテキストを「文書」と呼ぶ。

Brown コーパスでは、無作為抽出が 2 回行われているが、出版物から 2,000 語の文書を抽出した無作為抽出は、出版物の全てをデータとして採用するのが難しかったため、分量を少なくする目的で行われた無作為抽出であり、より重要なのは出版物リストから特定の出版物を選び出した 1 回目の無作為抽出である。これを行うに当たり、どのようなカテゴリーの出版物を何冊抽出するのかが前もって決断された。マケナリー・ハーディー（2014:361）では「代表性（representativeness）：各種のテキストタイプが現実の構成比と同等の比率で含まれるようにコーパスの標本が抽出されていること」と書かれていた。出版物リストから単純無作為抽出した場合、サンプルサイズが 500 では、「各種のテキストタイプが現実の構成比と同等の比率で含まれる」保証はない。そこで偶然の誤差によって偏った抽出が行われないように、あらかじめ各種のテキストタイプに一定の割合の出版物が含まれるように配慮したのである。

そのような配慮が行われたのは、テキストタイプが異なればそこで使用される言葉遣

いも大きく異なることがあらかじめ予想されたからである。たとえば「喫緊」という単語であれば恋愛小説より新聞に出現しやすいであろうことは想像に難くないし、同じ新聞でも報道より論説の方に出やすいことが予想される。これとは逆に「失恋」という単語なら新聞より恋愛小説に出現する可能性の方が高いであろう。このようなテキストタイプによる出現の傾向性は「失恋」のような話題語だけでなく、「述べる」のような動詞でも、あるいは受動態のような文法要素でも生じることが考えられる。このため、特定のテキストタイプに偏ることがないようにあらかじめ配分を決め、同じテキストタイプの中でもさらに偏りが出ないように無作為抽出を行ったのである。

しかし、たとえば恋愛小説というテキストタイプからたまたま選ばれた書籍の中の、さらにたまたま選ばれた 2,000 語の文書の中に、「失恋」という単語が何語含まれるかまで、「現実の構成比と同等の比率で含まれる」ことが期待できるとは考えにくい。つまり Brown コーパスの無作為抽出法は、単語や文法項目が母集団と同じ比率で抽出できるような方法は取られていない。無作為抽出されているのは出版物であり、データとして集積されたのはその代用である 2,000 語の文書である。

均衡コーパスが代表性を持つのは、個体が母集団から無作為抽出されているからであった。Brown コーパスで無作為抽出されているのは出版物（の一部の文書）であって、その文書の中に書かれている単語や文法項目ではない。Brown コーパスの個体は、文書だと考えられる。

第 2.2 項　BCCWJ の設計と無作為抽出法

次に、BCCWJ で無作為抽出されている個体は何であるかを検討する。BCCWJ は、2011 年に国立国語研究所によって公開された総語数約 1 億語のコーパスである。正式名の「現代日本語書き言葉均衡コーパス」という名称からすると、1 億語全体が均衡コーパスであると思われやすいが、実際は非常に複雑な内部構造を持っており、均衡コーパスといえる部分は 1,500 万語程度だといわれている（田野村，2014:123）。

図 2.1 は、BCCWJ の内部構造をそのデータ量に応じて描いたグラフで、円全体が約 1 億語である。この中で、縦縞で描いた固定長のうち、OW（特定目的・白書）を除いた部分（黒いコアデータも含む部分）が、真に均衡であるといわれているデータである。なお、統合形式とは、固定長と可変長を統合し、重複を除いたデータであるため、固定長が存在しない特定目的のデータは可変長と呼ぶのが正確だが、図 2.1 では統合形式という名称で統一した。

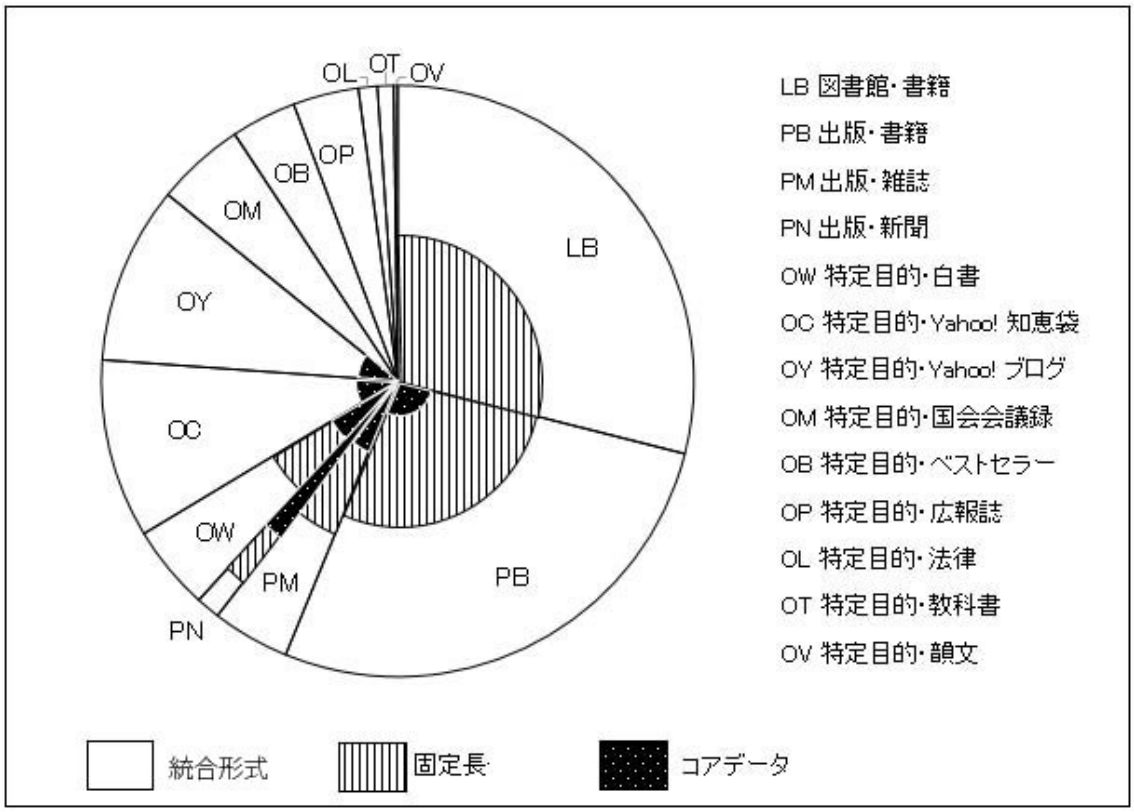


図 2.1 BCCWJ の内部構造とデータの割合

以下、国立国語研究所コーパス開発センター（2015）、および山崎誠（編）（2014）等を参考に、BCCWJ の設計の概略を記す。BCCWJ は大きく①図書館 SC、②出版 SC、③特定目的 SC の三つの SC に分けて設計されている。①図書館 SC は、書き言葉の流通実態を公立図書館の所蔵状態で近似的に把握することを目的として作られた SC で、都内の 13 の公立図書館に重複して所蔵されている 1986 年から 2005 年に発行された書籍を母集団とし、無作為抽出によって 10,551 文書、約 3,000 万語のデータが集積された。この中で文書の文字数を約 1,000 字に固定して抽出されたサンプルが固定長で、短単位で約 670 万語ある。個々の文書からは固定長とは別に、章や節などのある程度の文脈を確保した可変長サンプルが抽出された。可変長サンプルは概ね 1 万字を上限とするものの、字数はばらばらである。固定長と可変長の関係は、一部が重なっているものや可変長の中に固定長が含まれているものなどさまざまで、本研究では「固定長＋可変長－重複部分」を「統合形式」と呼んでいる。図 2.1 の統合形式はこれを指している。

コアデータとは、集積したデータを形態素解析する際に機械学習用に人手で修正を加えたデータで、解析精度が 99%以上あるといわれている（小木曽，2014:103-7）。形態素解析は格助詞の「で」と断定の助動詞の連用形の「で」など、区別が難しい言語項目

の解析精度は低く、区別が易しい言語項目の精度は 100%に近いなどのばらつきがあるが、コアデータ以外の平均的な解析精度は 98%だとされている（小木曽，2014:103-7）。

②出版 SC は、書き言葉の生産力という側面に着目して作られた SC で、2001 年から 2005 年にかけて出版された書籍（国会図書館に所蔵されている書籍）の母集団から約 2,800 万語強、雑誌（『雑誌新聞総カタログ』の 6 分類に入る雑誌）の母集団から約 440 万語、新聞（全国紙・ブロック紙・地方紙）の母集団から約 140 万語、合計約 3,400 万語のデータが集積された。この中で固定長サンプルは 850 万語である。

③特定目的 SC は、①、②の母集団には入らないが、現代日本語の書き言葉を研究する上で必要と思われる種類の書き言葉を収めた SC で、約 4,000 万語のデータが集積された。この中で白書だけは固定長サンプルがある。特定目的 SC は、公的な性格が強い書き言葉として OW（白書）、OT（教科書）、OP（広報紙）、OL（法律）、Web 上の書き言葉として OC（Yahoo！知恵袋）、OY（Yahoo！ブログ）、国会での発言を書き起こした OM（国会会議録）など、さまざまな側面からデータが集積されている。

以上のように BCCWJ にはレジスターと呼ばれている多様な媒体が存在し、それらの母集団はそれぞれに異なる。ここでは Brown コーパスの書籍と同じ媒体を対象にしている出版 SC 書籍レジスター（以下、出版書籍と略す）を例に、丸山、柏野（2014）を参照して具体的な無作為抽出の方法を確認する。

出版書籍の対象は、国立国会図書館に所蔵されている蔵書のうち、2001 年から 2005 年に発行された書籍である。ただし、漫画・写真集などの言語表現が主体でないものや、1 冊が 40 ページ以下の書籍などは除外された。Brown コーパスであれば、これらの書籍が除外された書籍リストから、乱数表によって特定の書籍が選択され、さらにその書籍から文書を抽出するページ数が選択される。しかし、BCCWJ ではこれとは異なる方法で文書が抽出されている。

まず、出版書籍の母集団を標本抽出の対象となる書籍の総文字数で定義し、48,539,925,351 文字とした。これは、さまざまな書籍の印刷面を合計で約 1,000 ページ調査し、1 ページ当たりの平均文字数を算出した上で、この平均文字数に発行された書籍の総ページ数をかけて推定された文字数である。母集団は発行年の 5 分類と「日本十進分類法（NDC）」のジャンル 11 分類で 55 層に層別された。出版書籍の母集団のページ数は 74,911,520 ページである。これを 55 層に分割し、各層ごとの全ページに対し、無作為に優先順位を割り振った上で、ページ内の 1 点を指定する座標情報を無作為に指定した。つまり、書籍を選んでそこから特定のページを選ぶという方法ではなく、「書

籍の総文字数」という母集団から、ダイレクトに 1 文字を無作為抽出する方法を取ったのである。この無作為抽出された 1 文字をデータの開始点、そこからちょうど 1,000 字目を終了点という。

ただし、書籍が入手できたものの、開始点を特定するはずのページが白紙などであった場合、本来ならこの書籍を放棄してもう一度リストから任意の 1 点を選び直すことになる。しかし再度別の書籍を入手するのはコストがかかるため、現実的には最初に選ばれた書籍から、任意の 1 点を選び直すこととした。つまり、理念的には「書籍の総文字数」という母集団から、ダイレクトに 1 文字を無作為抽出する方法を取ったが、現実的には、Brown コーパスと同様に、まず書籍を選択し、その書籍から任意のページの任意の開始点を選択することが行われた。そして、固定長であれば開始点が含まれる文の文頭から、終了点が含まれる文の文末までの約 1,000 字、可変長であれば開始点の 1 文字を含む節や章などの 1 万字までの構造的なまとまりを抽出したのである。

この無作為抽出法は、非常に厳密な無作為抽出を行っているようであるが、実は何を行っているのかの評価が難しい。前節で確認したように、無作為抽出の統計学的な定義は「有限母集団に対して，どの個体も等確率で取り出されるような標本抽出」のことである（尾畑，2014:160）。母集団を「書籍の総文字数」で定義したということは、現代日本語の書き言葉の個体を文字と考えたということである。そしてこの文字のリストが作られ、そこから一つの文字が無作為抽出された。出版書籍のサンプルサイズは 10,117 であるから、10,117 文字は 48,539,925,351 文字の母集団から確かに単純無作為抽出されている。この 10,117 文字を使用すれば母集団におけるひらがなと漢字の比率などが正しく推定できるであろう。しかし、コーパス言語学の関心は文字だけではない。このため単語の係り受けの関係などが分かるように、開始点を含む文の文頭と、終了点を含む文の文末までの約 1,000 字の文書を固定長として抽出した。固定長の場合、サンプルの開始点の 1 文字は無作為抽出されているが、残りの文字の抽出法についてはどのように考えればよいのであろうか。これについて丸山・柏野（2014:25-6）では次のように述べられている。

> 次に問題となるのは，抽出単位の決定，すなわち，個々のサンプルサイズをどの程度の大きさにするかという点である．これは，当該のコーパスを使ってどのような研究を実施するか，という使用目的とも密接に関連する問題である．
>
> たとえば，コーパスから得られる重要な知見の 1 つに，語彙頻度表がある．

BCCWJ の設計段階においても，語彙頻度表の作成が研究成果の 1 つとして想定されていた．仮に，母集団から無作為に 1 語ずつ抽出し，それを 1 億語分集めれば，母集団の特徴を十分に反映する語彙頻度表が完成することになる．しかしながら，そのような抽出は極めて手間がかかるうえ，収集したコーパスを語彙頻度表以外の用途に使えず，汎用的な目的が達成できない．語彙頻度表以外の研究目的，すなわち，語や句の意味の研究，文法研究，談話研究などにとっては，ある程度の文脈が確保されていることが必要となる．

逆に，より大きい範囲を抽出単位として採用すると，抽出したサンプルの中身が文脈による偏りの影響を大きく受ける可能性が出てくる．たとえば，1 冊の書籍をまるごと抽出単位とすると，サンプリング作業の負担は減るものの，たまたまその書籍に頻出していた語が大量に収録され，語彙頻度表の順位に影響する可能性がある．これでは，BCCWJ が備えるべき代表性という点に問題が生じることになる．（丸山・柏野，2014:25-6）

尾畑（2014:1）では、「統計的な調査の対象を一般に**母集団**と呼び，それを構成する各要素を**個体**と呼ぶ」とされていた。尾畑（2014:1）のいう「統計的な調査の対象」とは、丸山・柏野（2014:25）では研究目的の対象である語や句、文法、談話である。したがって研究目的によって母集団は、語、句、文法、談話の集合となり、それを構成する要素が無作為抽出されている必要がある。調査の対象が語の場合、語の集合が母集団で、そこから無作為抽出された語の集合が標本である。丸山・柏野（2014:25-6）が述べている「仮に，母集団から無作為に 1 語ずつ抽出し，それを 1 億語分集めれば，母集団の特徴を十分に反映する語彙頻度表が完成することになる」という標本抽出こそが厳密な意味での無作為抽出で、単純無作為抽出法とか個別抽出法などと呼ばれる。

しかし、そのような抽出法はコストが高く、汎用性もないため、一定量の文書を抽出したという。このような無作為抽出法は集落抽出法（cluster sampling）と呼ばれる。以下の引用は、小田（2009:179-80）による集落抽出法の定義である。

集落抽出法はクラスター・サンプリングとも呼ばれるように、何らかの塊（クラスター）を抽出単位として、抽出されたクラスター（集落）の構成要素全部を標本にする抽出方法である。集落をグループや群と読み替えてもかまわない。〔･･･〕。たとえば、ある市を対象に調査を計画したが、標本抽出に必要な世

> 帯や世帯員に関する情報が利用できなかったとする。そのときに、字や大字、丁目など何らかの区画（クラスター）を第一次抽出基準単位にして、その中から幾つかの区画を無作為に抽出し、抽出された区画の世帯すべてを調査する。
>
> 〔･･･〕。集落抽出法は、単純無作為抽出法よりも標本誤差が大きい。しかし、調査コストを軽減でき、詳細な抽出用リストが入手／利用できないときでも標本抽出が可能であるところは集落抽出法の大きな利点である。
>
> （小田，2009:179-80）

出版書籍・固定長の場合、無作為抽出されたある文字を開始点とし、それを含む文頭から約 1,000 文字を抽出単位にして、塊である文書が抽出されている。文書の構成要素は、文字に限定されるわけではなく、視点を変えれば語、文、談話とも考えられる。これらの集合によって文書ができあがっていると見なすことができるからである。ただし受動態などの文法要素や句などを集めても必ずしも文書の全体にはならないため、受動態や句などを調査対象にする場合はこれらを含む文などを、母集団の構成要素と見なすことになると思われる。談話も Brown コーパスや BCCWJ の固定長のように、語数や字数を一律に区切ったコーパスでは、談話の途中で抽出が打ち切られている可能性があるが、BCCWJ の可変長や書籍の全体を収録したコーパスなどでは、談話も構成要素と考えることが可能かもしれない。

ここでいう構成要素とは、これまで個体と呼んできたものと同じである。これらの個体の抽出用リストを作るためには、語であれば出版書籍の母集団の全書籍を形態素解析してリスト化しなければならず、事実上不可能である。この事情は文や談話でも同じである。しかし、文字の場合は、あくまでも推定ではあるが、リスト化することができた。そのリストから無作為抽出した文字を開始点とし、文字、語、文、談話を個体として集落抽出したと考えることが可能である。

前項では、Brown コーパスは出版物リストから任意の出版物を無作為抽出し、そこから 2,000 語の文書を抜粋しているため、この文書が Brown コーパスの個体であると述べた。しかし、上記のように考えると、Brown コーパスでも、文字、語、文が個体だと考えることができる。書籍から抽出された文書は、単純無作為抽出によって抽出された個体だが、その文書に含まれる文字、語、文も集落抽出法という無作為抽出法によって抽出された個体だという論理である。このように考えられるからこそ、現在のコーパス言語学では、文書ではなくこれらの言語単位を観察単位とした分析が行われていると考え

られる。

文字、語、文などの言語単位が無作為抽出された個体であると考えた場合、問題はこの集落抽出法によってこれらの個体がどれぐらいの標本誤差を持つかにかかっている。丸山・柏野（2014:26）でも、「より大きい範囲を抽出単位として採用すると，抽出したサンプルの中身が文脈による偏りの影響を大きく受ける可能性が出てくる」と述べられていた。集落抽出法の場合、「文脈による偏りの影響」によって、もはや無作為抽出とは呼べないほどの影響を受けるのであれば、文字、語、文などを個体と考えて統計分析を行う意義は薄い。そこで次節ではこの問題を考察する。

第 3 節　無作為抽出された個体は何か

本節では、はじめに木村（1982）、Baroni & Evert（2009）の研究を概観し、集落抽出法によって抽出された単語の頻度は、大きな標本誤差を持つことを確認する（第 3.1 項）。次に、Kilgarriff（2005）、Evert（2006）の主張を踏まえ、コーパスで無作為抽出された個体は文書だと考える方が妥当であることを述べる。最後に、文書を観察単位とした統計分析は、実践例が少ないため、体系的な方法論を検討し、分析法の実例を提案していく研究が必要であることを述べる（第 3.3 項）。

第 3.1 項　コーパス構築における集落抽出法の問題点

言語データを集落抽出法で抽出した場合、文脈による偏りの影響を受ける。この影響の問題を実証的に検証した研究に木村（1982）がある。これは、国定国語教科書の総索引を作る目的で作成された約 3 万 2 千枚のカードを使用して行われた。表 2.2 は、そこから 100 語を個別抽出（単純無作為抽出）した標本 10 個と、100 語を集落抽出した標本 10 個で、各標本ごとの異なり語数を比較した調査結果である。

表 2.2　抽出法による標本異なり語数の比較（n = 100）

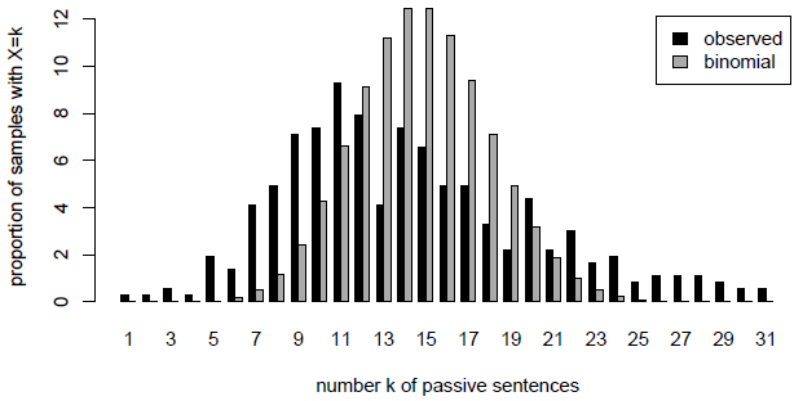
木村（1982:236）表 2 を引用

抽出法＼標本	1	2	3	4	5	6	7	8	9	10	平均	標準偏差	合併
個別抽出	70	67	71	68	73	66	67	65	70	69	68.6	2.33	436
集落抽出	71	48	56	56	56	57	61	46	56	61	56.8	6.58	379

異なり語数とは種類の異なる単語の数で、たとえば「標本｜を｜一つ｜一つ｜数える」

という文を例にすると、単語数の合計（延べ語数）は 5 だが、「一つ」は 2 回使われているため、種類が異なる単語の数は 4 である。このように種類の異なる単語の数を異なり語数と呼ぶ。表 2.2 を見ると個別抽出の平均は 68.6、集落抽出の平均は 56.8 で、集落抽出の異なり語数は、個別抽出より 17.2%ほど少ない。集落抽出では、ある話題に関して意味のある文脈が連なっているため、その話題に関する単語は何度も使用される。一方、個別抽出では、前後の文脈とは無関係に、それぞれが離れた場所から 1 語が独立して選択されるため、助詞や助動詞の機能語以外は、同じ単語が何度も重なって出現することはまれである。このような簡単な調査でも、集落抽出された文書の中の単語は、もはや無作為抽出されているとはいえないほどの誤差を持つことが分かる。

次に Baroni & Evert（2009）で、受動態がランダムに使用されているかどうかを調査した分析を概観する。ここでは Brown コーパスを使用し、各テキストに出現した受動態の頻度が二項分布に従うかどうかの分析が行われている。Brown コーパスの現実母集団は図書館の蔵書目録などをもとに作られた出版リストに記載されている出版物である。仮にこの出版物を全ての文に分解し、そこから 50 文ずつを単純無作為抽出した個体を 500 個作ったとする。この時、受動態がランダムに使われているとすれば、各個体に受動態の文が何文出現するかの出現確率は、二項分布に従うと考えられる。しかし、現実の Brown コーパスは 2,000 語単位で集落抽出されているため、各文書ごとに一定の傾向性を伴って出現する。実際の Brown コーパスの 500 の文書に出現した受動態の度数分布と、全体から単純無作為抽出された場合の理論的な二項分布を比較すれば、集落抽出によってどれだけの誤差が生じているかが分かる。この結果が図 2.2 である。



Baroni・Evert（2009:798）より引用

図 2.2　Brown コーパスにおける受動態文数ごとの相対文書度数と二項分布との比較（n = 500）

この分析では受動態を数える単位を文にしている。Brown コーパスの 500 の文書では、文の数が文書ごとに約 50〜200 と幅があるため、全ての文書から 50 文が再サンプリングされ、その文に受動態が何文含まれているかが調査されている。[3]

図 2.2 の横軸は、それぞれの文書から 50 文を再サンプリングした個体の中に、受動態が含まれている文の数である。一番少ない文書は受動態が 1 文、一番多い文書は 31 文である。縦軸は文書数の相対頻度で、これに 500 をかけると実際に出現した文書の数になる。実際の観測値が黒の棒、理論的な二項分布がグレーの棒である。

これを見ると、実際の観測値では二項分布より少なく出現している文書が多い一方で、二項分布の上限を超えるような文数で受動態が出現する文書も存在していることが分かる。1 文書 50 文当たり 30 文に受動態が使われているということは、文書の約 6 割で受動態が使われていることになる。集落抽出を行った場合、一つ一つの書籍の個性が色濃く反映され、単純無作為抽出を仮定した二項分布には従わない[4]。

以上、木村（1982）、Baroni & Evert（2009）で行われている研究結果を見ても、単純無作為抽出をした場合に比べて集落抽出をした単語や文の誤差は大きく、文書の中に含まれている単語や文を母集団から無作為抽出した個体と考えることには無理があることが分かる。

第 3.2 項　コーパスにおける独立した個体は何か

集落抽出された単語や文が無作為抽出とは言い難い分布をしている理由は、言語の本質的な性質に根差している。本項では言語の非ランダム性を指摘した Kilgarriff（2005）、Evert（2006）の主張を踏まえ、コーパスで無作為抽出された個体は文書であること、このため文書を観察単位とした研究が必要とされていることを述べる。

コーパスに含まれている単語や文は、無作為抽出された個体とは異なる分布を示す。

---

[3] これと類似の調査が Evert（2006）でも行われているが、そこでは各 2,000 語のテキストに含まれる受動態の頻度がそのまま使用されていた。受動態という言語現象を、語を基準にカウントするのか、文を基準にカウントするのかは難しい問題である。

[4] 言語研究の目的の一つはこのような特異な分布を示す文書はどのような属性を持っているかを明らかにするところにある。特異な文書の書名や文章の内容を調べた結果、全てがフォーマルな学術散文であることなどが判明するなら、受動態がどのような場面で多用されるかが分かる。これは第 1 章の「喫緊」がどのような文書で使われやすいのかという問いと同じである。コーパス言語学の研究では文書を観察単位にして図 2.2 のような度数分布図が作られることはほとんどないが、この図一つをとっても文書を個体として分析する有益性が分かる。

この理由は、言語の本質に根差している。すなわち、あまりにも当然のことではあるが、単語をランダムに使用して話をしたり、文を書いたりする人間はいない。誰しもが文法規則に則り、何らかの話題に沿って話をしたり文章を書いたりしている。そうなると、一度出現した話題語が何度も繰り返されたり、日本語であれば名詞の直後に助詞が続きやすかったりするのは当然である。

第 1 章でも述べたように、ある事象が起きる確率が他の事象が起きる確率に影響を与えている場合を統計の用語では「従属」といい、これと逆の場合を「独立」という。独立とは任意の i 番目と j 番目のデータに関して「j 番目の分布が i 番目の値に影響されない」ということである（豊田（編著），2009:26）。データに独立性がないと、統計分析の結果には深刻な影響が出る（栗田，1996 ; 森・吉田，2013:67-8 ; 豊田（編著），2009:26 ; 吉田，2001:248-50）。統計的検定は、独立性が確保されていることを前提として行われている。

この独立性はランダム性とも呼ばれる。先行研究では、言語はランダム性の仮定に違反しているため、統計的検定を行っても無意味だという主張がしばしば行われてきた。次の引用はカイ二乗検定が多くの場合で有意になってしまう問題を数多く論じてきた Kilgarriff の代表的主張である。

> Language users never choose words randomly, and language is essentially non-random. Statistical hypothesis testing uses a null hypothesis, which posits randomness. Hence, when we look at linguistic phenomena in corpora, the null hypothesis will never be true. （筆者訳：言語の使用者は決して単語をランダムに選ぶことはなく、言語は基本的にランダムではない。 統計的仮説検定は、ランダム性を仮定する帰無仮説を用いる。したがって、コーパスにおける言語現象を見る場合、帰無仮説は決して真ではないだろう。）
>
> （Kilgarriff, 2005:263）

これと同様の主張は Evert（2006:177）でもなされている。

> Words are chosen to convey a specific meaning or intention, and their arrangement follows the intricate rules of syntax. Therefore, we may ask: what has randomness to do with linguistics in the first place? And if there is nothing random about language, why should we apply statistical methods (based on the random sample model) at all?

（筆者訳：単語は特定の意味や意図を伝えるために選択され、その配置は複雑な構文規則に従う。このため、次のような問いが起こるかもしれない：そもそも、言語学に関してランダム性と関係しているものが何かあるだろうか？ 言語に関してランダムなものがない場合、どうして（無作為標本モデルに基づく）統計的方法を適用する必要があるのだろうか？） （Evert, 2006:177）

これらの主張のとおり、言語はランダムには使用されていないため、集落抽出法で抽出された文字、単語、文などを個体として統計分析を行う意義は薄い。最も自然なのは、Brown コーパスや BCCWJ の設計で実質的に行われていたように、文書を個体として統計分析を行うことである（Evert，2006；Baroni＆Evert，2009:796-9）。文書であれば、ある文書が選ばれたからといって、次の文書に何が選ばれるかには何の影響も与えていない。コーパスにおける独立した個体は文書である。推測統計を行うなら、独立した個体である文書を観察単位として分析を行う以外、正確な分析を行う方法はないと考えられる。

第 3.3 項　文書を観察単位とした分析法を体系化する必要性

コーパスにおける独立した個体が文書であるなら、文書を観察単位として統計分析を行うのは、ごく当たり前のことである。しかし、それほど当然のことがなぜこれまで行われてこなかったのであろうか。それを伺わせるような記述が Baroni & Evert（2009:797-8）で述べられている。

> Seeing how non-randomness effects can lead to a drastic misinterpretation of the observed frequency data, a question arises naturally: How can we make sure that a corpus study is not affected by non-randomness? While for many practical purposes it might be possible to ignore the issue, the only way to be absolutely sure is to ascertain that the unit of sampling coincides with the unit of measurement.〔･･･〕.
>
> This approach is only viable for phenomena, such as passive voice, that have a reasonably large number of occurrences in each text. It would not be sensible to count the proportion of occurrences of the collocation strong tea in the Brown texts (or even in a corpus made of larger text stretches), since the vast majority of texts would yield a proportion of 0% (in the Brown corpus, strong tea occurs exactly once, which means

that in all texts but one the proportion will indeed be 0%). （筆者訳：非ランダム性の影響が観測頻度データの大きな誤解につながるのを見ると、コーパス研究が非ランダム性の影響を受けないようにするにはどうすればよいのかという疑問が自然にわき起こる。多くの実用的な目的のために、この問題を無視することは可能かもしれないが、絶対的に確実な唯一の方法は、サンプリング単位が測定単位と一致することを確認することである〔･･･〕。

このアプローチは、受動態のように、各テキストにある程度出現するような現象に対してのみ実行可能である。Brown コーパス（またはより大きく拡大されたテキストで作られたコーパスでさえ）「strong tea」というコロケーションの出現割合を数えることは賢明ではないだろう。というのも、テキストの大部分の使用率は 0%になるであろうからである。（Brown コーパスでは、「storong tea」は正確に 1 回だけ出現する。これは一つのテキストを除く全てのテキストで、使用率が実際に 0%になることを意味している）。

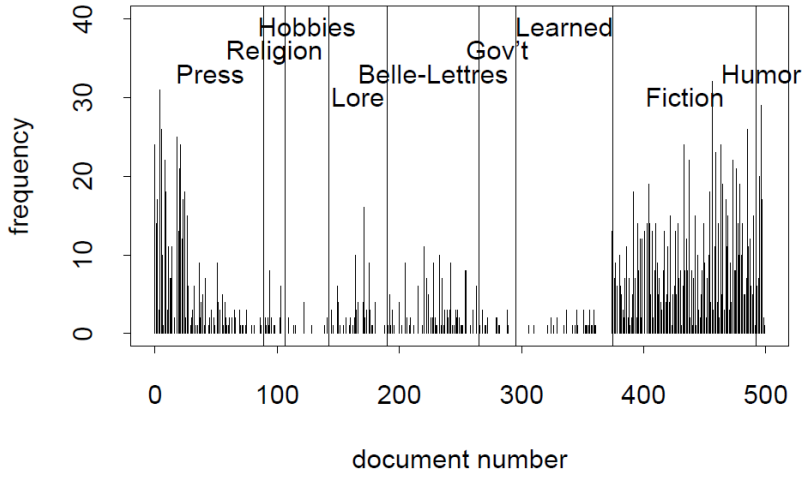
（Baroni & Evert，2009:797-8）

ここでは、ランダム性の仮定を満たす唯一の方法は、観察単位をコーパスのサンプリング単位に合わせることだと明確に主張されている。コーパスのサンプリング単位とは Brown コーパスでも BCCWJ でも実質的には文書であった。

しかし、ここで書かれている “for many practical purposes it might be possible to ignore the issue”（筆者訳：多くの実用的な目的のために、この問題を無視することは可能かもしれないが）とは何を意味しているのであろうか。統計分析の目的は、標本を使用して母集団の性質を推測することにある。単語が単純無作為抽出されているなら、それを使用して母集団の性質を推測することは容易だ。しかし、実際は集落抽出されているため、一つの文書に同じ単語が何度も出現したり、他の単語の出現に影響されて出現したりする。このような影響を無視できるとしたら、一つの文書にほぼ 1 回しか出現しない単語を使用して分析すること以外にはないのではないだろうか。しかし、どの単語が 1 文書に 1 回だけ出現するかは、文書を観察単位とした調査を行ってはじめて分かることである。非常に高頻度に出現する機能語であれば、集落抽出法の影響を受けないことも考えられるが、これも実際に調べてみるまでは何ともいえない。

分析の対象を話題語以外にしたからといって、集落抽出の影響が避けられるとは限らない。図 2.3 は、Brown コーパスの 500 の文書に「said」という単語が何回出現したか

を調査した図である（Church, 2000:182）。横軸は 500 の文書の数で、Press（新聞）、Religion（宗教）、Hobbies（趣味）などのジャンルごとに区分されて並んでいる。縦軸は一つの文書に出現した「said」の頻度である。



Church（2000:182） ”said” in Brown Corpus より引用

図 2.3 Brown コーパスに出現する「said」の頻度

これを見ると、Hobbies（趣味）や Learned（学術）には「said」が数回しか出現しないが、Press（新聞）の一部や、Fiction（小説）では、何回も出現していることが分かる。「said」は話題語ではないが、集落抽出の影響を受けている。何が集落抽出の影響を無視できる単語や言語的特徴なのかは、図 2.2 や図 2.3 のような、文書ごとに頻度を集約した分析を行わないかぎり分からないと思われる。

“This approach is only viable for phenomena, such as passive voice, that have a reasonably large number of occurrences in each text.”（筆者訳 : このアプローチは、受動態のように、各テキストにある程度出現するような現象に対してのみ実行可能である）という指摘は、第 4 章で行う検証結果とも一致しており、確かに正しい指摘である。文字、単語、文などの言語単位なら、文書の数に比べて個体数が大きく増えるため、もっと稀な現象でも統計分析できるように思える。しかし、いくら個体数が増えるからとはいっても、統計分析の前提を満たしていない分析を行う意義がどれほどあるかは疑問である。また、近年、コーパスの規模は飛躍的に増大しており、数百億語規模の Web コーパスも出現する時代になっている。統計的に正しいかどうか分からない分析を行うより、確実な分析を積み重ね、コーパス規模の拡大に伴って稀な言語現象を扱う方が、建設的ではないだろうか。

図 2.3 では、「said」が全ての文書に出現しているわけではないが、どのジャンルによく出現し、どのジャンルに出現しないのかという興味深い現象を観察することができる。どんな文書には出現しないのかも、有益な情報の一つである。また、グラフの縦棒は具体的な文書に対応しているため、「said」が頻出する文書の書名や実際のテキストを確かめることもできる。何らかの言語的特徴を調べたとき、出現数がこれよりもっと少なかったとしても、図 2.3 と同じように、ジャンルによる偏りなどが発見できるのであれば、有効な分析になると思われる。

Baroni & Evert（2009:797-8）では、コーパス分析においてランダム性の仮定を満たす唯一の方法は、観察単位をコーパスのサンプリング単位に合わせることであると明確に主張されていた。しかし、文書を観察単位とした研究の意義や具体的な分析方法が、十分に明らかにされているわけではない。

Baroni & Evert（2009:797-8）の主張の根拠となった Brown コーパスにおける受動態の分析においても、その前身の研究である Evert（2006）では、各文書における受動態の頻度がそのまま使用されていた。しかし、Baroni & Evert（2009:797-8）では、一つの文書から 50 文を無作為抽出した中で受動態が使用されている文の数が観測値として使用されている。研究の主旨は Brown コーパスの文書における受動態の使用の分布は、二項分布には従わないということだが、分析の方法が異なっている。この例を見ても分かるように、一口に「文書を観察単位とした分析」といっても、具体的にどのような分析を行えば有効な分析になるかについては、さまざまな研究テーマごとに、分析例を積み上げながら検討していく必要がある。

集落抽出法という考え方に基づけば、コーパスで無作為抽出されている個体は文字、単語、文などの言語単位であると考えることが可能に思える。しかし、これらの言語単位は、「ある町内に住んでいる人間」のように、互いが独立して存在しているわけではない。一つの文書内の言語単位は、互いに密接な関連性を持って出現している。これは統計分析の前提となる独立性の仮定に違反するため、言語単位をコーパス内の独立した個体と考えることは難しい。一方、出版物の一部である文書は、ある文書が選ばれたからといって、次の文書の選択には何の影響も与えない。コーパスで無作為抽出された個体は文書であり、コーパスは文書を個体として構築されていると考えるのが妥当である。しかし、文書を観察単位とした分析法は、これまでほとんど使用されてこなかった。実際にどのような方法を用いれば有効な分析ができるかについて、十分に明らかにされているとは言い難い。このため、文書を観察単位とした体系的な分析法を検討していくこ

とが求められる。本研究の必要性はまさにこの点にある。

第 4 節　学習者コーパスにおける個体は何か

本節では、広義コーパスの一つである学習者コーパスを取り上げ、学習者コーパスにおける個体とは何かについて考察する。第 4.1 項では、現在公開されている主な日本語学習者コーパスを概観し、これまで学習者コーパスを使用してどのような研究がなされてきたのかを簡単に振り返る。第 4.2 項では、学習者を観察単位とした Woods, Fletcher, & Hughes（1986）の分析法と、言語単位を観察単位とした松田・宮永・庵（2013）の分析法とを比較し、学習者コーパスにおける個体は言語単位ではなく、学習者と考えるのが妥当であることを述べる。

第 4.1 項　日本語学習者コーパスの概観

学習者コーパスの歴史は浅く、1990 年代初頭に始まる（グレンジャー（編）, 2008:4）。学習者コーパスの研究が最も進んでいるのは英語教育の分野であり、英語学習者コーパスで最大規模のコーパスは Cambridge Learner Corpus の 4,500 万語といわれる（投野・金子・杉浦・和泉，2013:10）。ただし、このコーパスは辞書作成などの商用利用を目的に構築されたもので、非公開である。一般公開されているものでは、15 の母語別に 20 万語（中国語のみ 50 万語）のデータを集めた International Corpus of Learner English（ICLE）の約 370 万語が最大であるという（石川，2012:221）。日本語学習者コーパスの規模はこれよりはるかに小さなものが多かったが、現在、国立国語研究所で構築されている I-JAS は、最終的に 340 万語程度の規模になるものと予想される（I-JAS の設計については第 5 章で紹介する）。

学習者コーパスは、学習者の発話を録音し、文字化した発話コーパスと、学習者が書いた作文を収集した作文コーパスの 2 種類に大別される。表 2.3 は、望月（2012:115）「表２．公開されている外国人日本語学習者コーパス」、および、迫田・小西・佐々木・須賀ほか（2016:95）「表１　主な日本語学習者発話コーパス」を参考にしながら、現在公開されなくなったコーパスを除くなどの筆者の判断で取捨を行い、主な日本語学習者コーパスの概要をまとめた表である。ただし、これ以外にも公開されているコーパスがあるため、この表ですべての公開コーパスが網羅されているわけではない。

表 2.3 の項目で、製作年に「初版」という限定を行ったのは、ここに記載した年の後で、データが追加されたコーパスが存在するためである。また、I-JAS は現在製作中で

2020 年に完成予定である。C-JAS（Corpus of Japanese As a Second language）の場合、製作年という点からいうと広く公開されたのは 2014 年になるが、データが取得された最終年は 1994 年であるため、この年を記入している。言語研究を行う場合、日本語教育を取り巻く環境や日本語そのものが、かなり早いサイクルで変化している。コーパスを分析に使用する場合はそのデータがいつ頃取得されたかのという情報が必要だと考え、このような措置を取った。

表 2.3　主な日本語学習者コーパスの概要

種別	名称	中心的製作者・団体	初版製作年	データ量	学習者の母語・人数	データ収集方法
会話	Corpus of Japanese as a second language：C-JAS	迫田久美子 国立国語研究所	1994	47本 57万語?	中国語3名・韓国語3名	NSとの3年半の自由会話
会話	KYコーパス	鎌田修・山内博之	1999	90本 17万語	中国語30名・韓国語30名・英語30名	OPI
会話	日本語学習者による，日本語・母語対照データベース：発話対照DB	宇佐美洋 国立国語研究所	2004	190名	中国語69名・韓国語70名・タイ語51名	朗読・スピーチ・ロールプレイ
会話	日本語学習者会話データベース	野山広 国立国語研究所	2007	339本	韓国語170名・中国語58名・英語28名・その他83名	OPI
会話	BTSJによる日本語話し言葉コーパス	宇佐美まゆみ	2007	294会話 66時間	日本語・韓国語・中国語・英語・その他	対話：雑談・電話・論文指導
会話 作文	International Corpus of Japanese as a Second Language：I-JAS	迫田久美子 国立国語研究所	2016	最終1,050人 推計340万語	中国語200名・韓国語100名・英語100名・その他	インタビュー・ロールプレーなど8種類
作文	日本語学習者による，日本語・母語対照データベース：作文対訳DB	宇佐美洋 国立国語研究所	2000	1,500編 300～800字	中国語、韓国語、モンゴル語、、タイ語、ベトナム語等20か国語	作文と執筆者本人による母語訳
作文	日本語学習者作文コーパス	李在鎬・伊集院郁子 2010科研グループ	2012	304名 11.4万語	中国語144名・韓国語160名	2つのテーマによる作文
作文	学習者作文コーパス「なたね」	仁科喜久子 「ひのき」プロジェクト	2012	285編 20.5万文字	中国語115名・マラティー語36名・ベトナム語13名・他28名	作文に誤用タグ添付
作文	日本語教育のためのタスク別書き言葉コーパス：YNU書き言葉コーパス	金澤裕之	2014	1080編 25万語	中国語30名・韓国語30名・日本語30名	12種類のテーマによる作文

これらのデータ量で、語数を書いているものは、基本的にコーパスが形態素解析され、その語数が紹介されているものである。しかし、学習者ごとの語数が一覧表の形で詳細に公表されているものは少なく、語数が明確に特定できないコーパスもある。たとえば、KY コーパスの語数はこれを形態素解析して公開している「タグ付き KY コーパス」の語数を参照したが、李・淺尾・濱野・佐野ほか（2008）では 173,198 形態素、李（2009）では 232,605 語と記されている。筆者が「タグ付き KY コーパス」を使用して学習者の全品詞をダウンロードして集計した形態素数は 170,454 であった。李在鎬氏によればこのような違いが起きる原因はその都度修正しながら作業しているためで、現時点でダウンロードできる形態素数を総語数と考えてよいとのことである（2017.07.15 のメールによる）。このため筆者がダウンロードした語数を現時点の語数であると考えることにした。

C-JAS の 57 万語は、コーパスの紹介サイトの記事に従ったが[5]、ここでは、「【データ量】47 本（計約 46 時間 30 分　約 57 万語）」とあるだけで、これが学習者のみのデータ量なのか、日本語母語話者のデータも含めた値なのか明記されていない。一方、KY コーパスのデータ量は学習者のみの語数である。この収録時間を 90 本×25 分＝37.5 時間と仮定すると（鎌田，2006:43）、37.5 時間で 17 万語であるため、C-JAS の 46 時間 30 分で 57 万語という語数には日本語母語話者のデータも含まれていると考えるのが妥当であろう。よって、学習者のデータは、このおよそ半分の 29 万語程度かと思われる。同様に KY コーパスの語数を目安にしてごく大まかに計算すると、日本語学習者会話データベースの学習者の語数は約 64 万語、BTSJ による日本語話し言葉コーパス（以下、BTSJ と略す）の語数は 30 万語程度と推定される。ただし、BTSJ の場合、全てが母語話者と学習者の会話ではないため、学習者のみの発話語数は推定が難しい。

作文コーパスではデータ量が文字数で示されているものがある。BCCWJ 図書館 SC の固定長は約 1,000 字で、約 635 語（短単位）であるから、これを利用して概算すると、学習者作文コーパス「なたね」の語数は約 13 万語、作文対訳 DB の字数を 500 字×1,500 編＝75,000 字と仮定すると、約 48 万語程度と思われる。ただし、学習者の作文ではひらがなが多く使用される傾向があるため、語数はこれより少なくなることが予想される。なお、YNU 書き言葉コーパスは製作者による形態素解析がなされていないため、これを分析した趙（2015:295）の値を記している。

これらのコーパスを見ると、100 万語を超えるコーパスは、I-JAS のみである。石川（2017:69）では、学習者コーパスの量的な課題に関して、「学習者属性（性別・習熟度・動機づけのタイプ等）を統制した分析を行うにはデータ量が不足していること（主要な母語話者コーパスが 1～4 億語であるのに対し、学習者コーパスは 100～400 万語程度）等は、残された課題と言えるでしょう」と指摘している。これは、100 万語のコーパスであっても、学習者の習熟度別に 4 段階に分けた場合、各レベルは 25 万語程度になり、それでは少なすぎるということを意味している。日本語学習者コーパスでは、全体で 25 万語程度のコーパスが少なくないため、その精度には注意が必要である。

学習者コーパスを使用する上では、量の問題とともに、どのような方法でデータが取得されているかも重要である。KY コーパスや日本語学習者会話データベースでは OPI（Oral Proficiency Interview）のデータが使用されている。このため、これらのコーパスでは OPI の実施方法に強く影響を受けていると考えられる（これについては、第 5

[5] http://lsaj.ninjal.ac.jp/?page_id=134（2018.12.15 閲覧）

章で詳述する)。また、発話対照 DB では、学習者の自発的な発話だけでなく、朗読のデータも混在しているため、分析には注意が必要である。作文コーパスでは、トピックが統制されているか、辞書を使っているかなどの条件により、データが影響を受けるため、これらの設計方針を確認した上で分析を行う必要がある。また、学習者の習得レベルや学習歴などの情報がどれぐらい備わっているかも、分析する上では重要である。

表 2.3 では I-JAS と YNU 書き言葉コーパス（以下 YNU と略す）が複数のタスクを行ってデータを取得した新しいタイプのコーパスで、I-JAS では 8 種類のタスク、YNU では 12 種類のタスクを行ってデータが取得されている。このため、特定のタスクに出現しやすい言語形式が学習者間で比較しやすいだけでなく、全てのタスクのデータを総合すると学習者の言語能力を過不足なく代表したデータになっていると考えられる。また、学習者の習得レベルも J-CAT（Japanese Computerized Adaptive Test)、SPOT（Simple Performance‐Oriented Test)、日本語能力試験など、複数のテストの結果が記載されており、学習期間や教育機関などの情報も調査されているため、どのような属性を持った学習者がどのような言語的特徴を持っているのかが調査しやすくなっている。

このような学習者コーパスを使用して行われる研究のテーマとして、Leech（2008:xi）では、(1）過剰使用や過少使用しがちな目標言語の言語的特徴、(2）母語転移、(3）回避ストラテジーを使用する言語領域、(4）母語話者的な運用や非母語話者的な運用がなされる言語領域、(5）非母語話者的な運用がなされる言語領域の頻度順（特別な助けを要する順番)、(6）上記の母語別比較の 6 点があげられている。また、日本語教育におけるコーパスの活用について啓蒙的に紹介した砂川（2011:14）では、「学習者コーパスを活用した研究は，誤用の研究を進めるだけでなく，非用や過少使用，レベルごとの習得状況を明らかにすることなどを可能にし，日本語教育や第二言語習得研究のさまざまな領域に有用な知見をもたらしてくれる」と述べられている。日本語教育で多用されてきた KY コーパスを使用した研究ではこの他に、学習者の習得状況から逆算してシラバスを構築する研究（山内，2009）なども行われている。問題は、これらの研究がどのような方法で行われているか、特に、何を観察単位として行われているかにある。

第 4.2 項　学習者コーパスにおける個体は何か

前節で概観したように、学習者コーパスでは学習者にインタビューを行ったり、作文を書いてもらったりしてデータを取得している。このため、学習者コーパスにおける個体は、学習者（が産出した会話や作文）だと考えられる。均衡コーパスでは、出版リス

トから出版物が無作為抽出され、その出版物の代表として一部の文書が抽出されていた。この構造を学習者コーパスに当てはめると、学習者コーパスを製作するに当たっては、まず学習者が選ばれ、その学習者の言語的特徴を反映するデータとして会話や作文などの産出物（文書）が収集される。学習者コーパスと均衡コーパスの違いは、学習者が無作為抽出されていない点と、一人の学習者から、複数の文書が収集される場合がある点である。一人の学習者から、一つの文書が収集されているコーパスの場合は、均衡コーパスと同様に文書を観察単位と考えてもよいが、一人から数種類の文書が収集された場合、それらの文書は同じ学習者の言語的特徴を持っており、互いに独立していない。このため、学習者コーパスでは、観察単位は文書ではなく、学習者と考えるのが妥当である。

このような人間単位で複数の文書を併合する利点は、先行研究でも確認されている。石田・佐藤（2010）は、30 人の職業作家の 90 冊のエッセイ集から抽出したエッセイコーパスを使用し、著者推定実験を行った研究である。この実験では 1 人当たり 5,000 字のテキストを使用した場合の推定精度が 97.8%、1 か所から 1,000 字を抽出したテキストを使用した場合は 74.4%、5 か所から抽出した 200 字を併合した 1,000 字を使用した場合は 84.9%となったと報告されている。この研究では、データ量が多いことが推定精度を高める最大の要因となっているが、同じ 1,000 字のデータであれば、一続きの文章から連続して抽出するより、複数の箇所から少しずつ抽出して併合した方が、精度が高くなっている。YNU や I-JAS のデータも複数のタスクを併合することにより、その学習者の言語的特徴をよりよく反映したデータになると考えられる。

均衡コーパスは、母集団から文書という個体を無作為抽出することで代表性を確保している。しかし、コーパスにはさまざまな種類が存在し、割合からいえば無作為抽出によって代表性が確保されていない広義コーパスの方が圧倒的に多い。学習者コーパスは、均衡コーパスの対極にある広義コーパスの一つである。学習者コーパスから得られた観測値は、母集団から無作為抽出されていないため、統計的検定を行っても、その結果が母集団に一般化できるわけではない。それにもかかわらず学習者コーパスを使用した計量的な分析で、統計的検定などが行われてきた理由は、検定の結果がその学習者コーパスに限定されるにしても、個体数の不足や比較している観測値の差が小さすぎるために、分析結果に意味が見いだせなくなっていないかどうかを検証するためだと思われる。問題は、このような統計的分析にどのような観測値を使用すべきなのか、すなわち、学習者コーパスの個体は何なのかである。

学習者コーパスができる以前から、学習者の作文や発話データなどを使用して分析を行う研究は多かった。Woods, Fletcher, & Hughes（1986）は学習者のデータを使用した統計分析の方法を解説した書籍で、これまで広く使用されてきた。ここでカイ二乗検定の説明に取り上げられているのは、アメリカ生まれの 14 歳の生徒（Group A）とメキシコ生まれでアメリカに在住している 14 歳の生徒（Group B）各 30 名における動詞時制の誤用の研究である[6]。この調査では、生徒に短い映像を見せ、それについてに書かせた 100 語以上の作文から、先頭の 100 語を抜き出した部分に出現した動詞過去形の誤用数を調査し、出生地の違いによって誤用の出現傾向に違いがあるかどうかを比較している。この検定の個体は学習者であり、学習者を誤用数 0、誤用数 1、誤用数 2-6 の区間で区分し、その度数（人数）を比較している。表 2.4 はこれを記したクロス表である。表の下に示した効果量（Cramer's V : クラメールの連関係数）は筆者が書き加えた。

表 2.4 メキシコ移民と母語話者の動詞時制の誤用数

Woods, Fletcher, & Hughes（1986:140）Table 9.4 より一部抜粋

表の下の効果量は、筆者が加えた

	0	1error	2-6 errors	Row total
Group A	7	7	16	30
Group B	13	11	6	30
Column tot	20	18	22	60

$\chi^2 (2) = 7.22, p < .05$, Cramer's $V = .347$

表 2.4 では、アメリカ生まれの生徒よりメキシコ生まれの生徒の方が、誤用を犯す生徒が少ないことが見て取れる。この分析は、規模は小さいものの、学習者 60 名による作文コーパスを使用した分析と考えることも可能である。学習者コーパスが作られる以前の分析は、このように学習者を観察単位にした分析が行われていた。しかし、学習者データが「コーパス」という名称で呼ばれるようになると、学習者数ではなく誤用数や正用数などの言語単位の頻度が分析に使われるようになる。これは均衡コーパスの分析で言語単位を観察単位とした研究が行われている分析法を見習ったためだと思われる。

表 2.4 の分割表に替えて、単語を観察単位として分割表を作り、カイ二乗検定を行うと表 2.5 のようになる。表 2.4 で「2-6 errors」の区間は、詳しい誤用数が分からないた

[6] Woods, Fletcher, & Hughes（1986）は言語研究のための統計分析の解説書であり、この研究自体は Ferris, M. R., Politzer, R. L.,（1981）”Effects of early and delayed second language acquisition: English composition skills ofSpanish-speaking junior high school students”, TESOL Quartely 15, pp.253-274. によってなされている。

め、この区間の生徒の誤用数はすべて平均の 4 と仮定してクロス表を作成した。生徒の作文は 100 語ずつ使われているため、「100 語－誤用数＝それ以外の語数」である。

表 2.5 メキシコ移民と母語話者の動詞時制の誤用数（観察単位：単語）

	誤用数	それ以外の語数	合計
Group A	71	2,929	3,000
	2.4%	97.6%	100.0%
Group B	35	2,965	3,000
	1.2%	98.8%	100.0%
合計	106	5,894	6,000

$\chi^2(1) = 11.764, p < .01$, Cramer's $V = .044$

表 2.5 は、二つのグループにおける各 3,000 語の作文の中に、動詞時制の誤用数が何語含まれているかを比較している。この表では、一人一人の学習者の性質はもはや問題にされておらず、各グループから集めた 3,000 語の言語データとそこに含まれている誤用数だけが問題にされている。単語を観察単位にしているため、使用した個体数は 6,000 という大きなサイズになる。これほど個体数が多いと、p 値はほとんどのケースで有意になる。しかし、二つのグループに実際にどれぐらいの違いがあるかを示す効果量の Cramer's V（クラメールの連関係数）は.044 と非常に小さく、二つのグループに実質的な違いは認められない。

学習者を観察単位とした表 2.4 の分析と、単語を観察単位とした表 2.5 の分析では、どちらが有効な分析になっているのだろうか。表 2.4 では、アメリカ生まれの学習者とメキシコ生まれの学習者をそれぞれ無作為抽出したわけではない。このため、この結果は調査を行った学校の生徒だけに限定される結果かも知れない。しかし、知りたかったことは、メキシコ移民の生徒が、必ずしも文法的な誤用を犯しやすいわけではないという全体的な傾向性であろう。全体的な傾向性といっても、生徒一人一人はさまざまな個性や能力の違いがあるため、これを一括して論じることはできない。そこで各グループを構成する個体を学習者と考え、一人一人の誤用数を観測値にして統計分析を行っている。学習者が A、B、C、・・・、と 30 名いる中で、学習者 A の誤用数と学習者 B の誤用数には何の関係もない。学習者同士は独立している。一方、表 2.5 の分析が問題にしているのは各グループの 3,000 語の単語である。その 3,000 語の単語を一つ一つ調べて、動詞の時制を誤っているかいないかを判別し、誤用の出現確率を比較するという枠組みで分析がなされている。個体は単語である。しかし、たとえば文法が得意な学習者 A

と文法が不得意な学習者 B がいた場合、学習者 A の誤用は少なく、学習者 B の誤用は多いことが考えられる。3,000 個の単語は互いに独立しているわけではなく、学習者ごとの 100 語のまとまりで強い関連性を持っている。これは、独立性の仮定に違反している。3,000 個の単語が、一つ一つ独立した個体として振る舞っていて、ある時は正用、ある時は誤用としての振る舞いをランダムに見せているわけではない。

学習者コーパスの例で考えると、学習者コーパスの個体が学習者であること、すなわち観察単位が学習者であることは当然のことに思われる。しかし、コーパス言語学では、文字、単語、文などの言語単位を観察単位とした分析が主流となっており、学習者のデータを使用した研究でも、コーパスという名称が一般的になるに従って、言語単位を観察単位とした分析が行われるようになっていく。

KY コーパスを使用した研究でも、その初期では学習者を観察単位とした研究が行われていた。田中（1999）、許（2000）、坪根（2002）などでは、いずれも学習者全員のデータが個別に調査され、それぞれの観測値を記載した表が掲載されている。分析もこれらの表を基にし、学習者の習得レベルの違いによって、調査対象の頻度が増加していく様子などが観察されている。ここではカイ二乗検定や t 検定などは行われてはいないが、これらは明らかに学習者を観察単位とした分析である。特に許（2000）や坪根（2002）では、使用しているデータを「KY コーパス」という名称ではなく、「OPI データ」という名称で呼んでいる点が象徴的である。これに対し、コーパスという名称が定着するに従って、このような学習者ごとに観測値を集約するという分析法は影を潜めていく。

松田・宮永・庵（2013）は、KY コーパスを使用し、文という言語単位を観察単位にして分析を行った研究である。次にこの研究の中でカイ二乗検定を行っている分析を取り上げ、学習者を観察単位とした分析との比較を行う。松田・宮永・庵（2013）は、KY コーパスの超級 9 名、上級 9 名のデータを使用し、超級を特徴づける談話特性がどのような表現であるかを調査している。KY コーパスの詳細は第 5 章で述べるが、OPI をもとに、初級から超級までの中国語、韓国語、英語母語話者計 90 名の音声データを書き起こしたコーパスである。インタビューは最長 30 分で、学習者の習得レベルを判断するためのロールプレイなども行われる。上級の発話語数平均は 2449.4 語、標準偏差 609.9 語、超級の発話語数平均は 2866.0 語、標準偏差 684.3 語で、超級の方が語数が多い。

松田・宮永・庵（2013）では詳細な分析に入る前に、超級者と上級者で「こ系」「そ系」「あ系」の使用傾向を比較するためのカイ二乗検定が行われている。「こ系」は「この前、このごろ、こう、こんなに、このー、こんなふうに、これから、これ」、「そ系」

は「そのまま、その後、その、そうですね、そんな、そのー、そんなに、そうね、そうすると、そう、それほど、そんなふう、それだけ、それ」、「あ系」は「あの、ああ、あんなに、あのー、あれ」という表現がコーディングに使用されている。集計単位は文で、これらの表現が出現した文数を使用して分析が行われている。その結果が表 2.6 である。

表 2.6 上級・超級におけるコ系、ソ系、ア系の語使用（KH-Coder）

松田・宮永・庵（2013）表 8 より引用

	こ系	そ系	あ系	ケース数（文）
超級	98（18.08%）	194（35.79%）	122（22.51%）	542
上級	33（3.52%）	237（25.27%）	123（13.11%）	938
合計	131（8.85%）	431（29.12%）	245（16.55%）	1480
カイ二乗値（自由度1）	88.504**	17.935**	21.280**	

**は有意水準 1%を指す。

表 2.7、表 2.8 は松田・宮永・庵（2013:51）の表 9 から「そ系」に関する部分を抜粋し、超級と上級に分けて筆者がまとめ直した。KY コーパスでは各母語ごとに、超級が 5 名（計 15 名）、上級が 10 名（計 30 名）存在するが、この分析では各母語の ID 順に、先頭から 3 名ずつが選ばれて使用されている。観測値は「そ系」が使用されている文の数である。右端の「文数」は、学習者の総発話に使用されている文の数、「割合」は総発話文数に占める「そ系」が出現した文数の割合である。

表 2.7 超級の「そ系」使用文数

松田・宮永・庵（2013:51）表 9 をもとに作成

超級ID	そ系	割合	文数
CS01	19	41.3%	46
CS02	29	51.8%	56
CS03	25	32.9%	76
ES01	23	26.4%	87
ES02	26	40.0%	65
ES05	21	45.7%	46
KS01	13	22.8%	57
KS03	14	24.1%	58
KS06	24	48.0%	50
合計	194	35.9%	541

表 2.8 上級の「そ系」使用文数

松田・宮永・庵（2013:51）表 9 をもとに作成

上級ID	そ系	割合	文数
CA01	33	22.3%	148
CA02	31	39.2%	79
CA03	29	22.8%	127
EA01	43	45.3%	95
EA02	21	18.1%	116
EA03	30	27.5%	109
KA01	21	26.9%	78
KA02	24	42.9%	56
KA03	5	3.9%	129
合計	237	25.3%	937

表 2.7 と表 2.8 の文数を比較すると、超級は 60 前後でばらつきが少ないのに対し、上

級は 56～148 とばらつきが大きい。超級は上級に比べて語数平均が多いため、超級の文数が少ないということは、それだけ 1 文当たりの語数が多いことを意味している。このことは、超級ほど長く複雑な文で発話していることを示唆している。上級でも超級の学習者と同じような発話傾向を持つ学習者も存在するが、短い文で発話する学習者も交じっており、そのような学習者は文数が多くなっている。また、上級では文数が多い学習者の「そ系」割合が低い傾向が見られる。文数が 100 を超える学習者の割合は、すべて 30%以下である。超級、上級を合わせた学習者 18 名の「そ系」割合の平均は 32.3%、標準偏差 12.3%である。

表 2.9 は松田・宮永・庵（2013:50）の表 8 から「そ系」に関する部分を抜粋し、クロス表を作成したものである。効果量は、筆者が書き加えた。この分析の観察単位は文になっている。一方、表 2.10 は、学習者を観察単位にして新たに筆者が作成した表である。「そ系」の割合が 30%未満か 30%以上かで区分した学習者数を使用してクロス表を描き、フィッシャーの正確確率検定を行った。カイ二乗検定ではなく、フィッシャーの正確確率検定を行ったのは、カイ二乗検定の場合、期待値が 5 未満になると、解析精度が下がるといわれているためである。

表 2.9 レベル別「そ系」比較（文数）

松田・宮永・庵（2013:50）表 8 をもとに作成

表の下の効果量は、筆者が加えた

	そ系	そ系以外	合計
超級	194	348	542
	35.8%	64.2%	100.0%
上級	237	701	938
	25.3%	74.7%	100.0%
合計	431	1049	1480

$\chi^2(1) = 17.93, p < .01$, Cramer's $V = .110$

表 2.10 レベル別「そ系」比較（学習者数）

	30%未満	30%以上	合計
超級	3	6	9
	33.3%	66.7%	100.0%
上級	6	3	9
	66.7%	33.3%	100.0%
合計	9	9	18

$p = .20$ 非有意, Cramer's $V = .333$

観察単位を文にした表 2.9 と学習者を観察単位とした表 2.10 では、どちらがより有効な分析になっているのだろうか。はじめに学習者を観察単位にした表 2.10 から検討する。「そ系」割合の平均が 32.3%であるため、表 2.10 では、学習者を「そ系」をあまり使わない学習者（「そ系」割合 30%未満）と、よく使う学習者（30%以上）に区分し、超級と上級でその分布に差があるかどうかを検定した。検定に使用した人数が 18 名しかいないため、p 値は非有意であるが、効果量の Cramer's V は .333 で、中程度の関連性を持っていると判断される。つまり、「そ系」は超級を特徴づける談話特性になって

いる可能性が高い。ただし、*p* 値が非有意であるため、分析対象の学習者を増やし、「そ系」割合の区分も 3 区分以上に増やすなどして再分析することが望ましい。

一方、表 2.9 は表 2.7 と表 2.8 における各レベル 9 名の文数の合計を使用して比較を行っている。分析の有効性を判断するポイントは文という言語単位が、コーパスにおける独立した個体であるかどうかである。上級学習者の表 2.8 で特に顕著な特徴が見られるように、学習者によって文数が多い傾向と、文数が少ない傾向がはっきり分かれている。文数が多い学習者の文は短く、文数が少ない学習者の文は長いと考えられる。文は学習者に従属しており、一つ一つが独立しているわけではない。表 2.9 は、一つ一つが独立している文を抽出した上で、その文に「そ系」が含まれているかいないかを分析するための枠組みを使用している。この分析で使われているデータにおいて、文は独立性の仮定に違反しているため、有効な分析になっているとは考えにくい。

また、学習者の OPI レベルごとに、文の数を合計することも問題である。たとえば、表 2.8 の KA03 の学習者は「そ系」を使用した文数が 5 になっている。この値は他の学習者に比べて著しく低い。この学習者が他の学習者と同じぐらい「そ系」を使用していたら、上級の合計数は 10%程度増えていたと思われる。合計数はこのように、一人の学習者の値によって左右される。しかし、「そ系」をたくさん使うか使わないかは、学習者個々人の問題であって、一人の学習者が極端に使わないからといって、そのレベル全体の値が低くなるのはおかしい。これはまるで学習者レベルで分けたグループが連帯責任を負っているようなものである。KA03 の学習者の使用数は、他の 8 名の使用数とは無関係のはずである。一方、学習者を個体と考えれば、このような影響を及ぼすことなく、そのレベルにどのような学習者がどれぐらいいるかという比較ができる。つまり、表 2.9 のように、学習者が産出した言語単位を観察単位とするのは、有効な分析になっていないと考えられる。

以上、学習者コーパスにおける個体とは何かを検討してきた。学習者のデータを使用して統計分析を行うのであれば、学習者を観察単位にするのはごく当たり前だと思われる。実際、学習者コーパスが出現する以前の研究では、そのような分析がなされていた。しかし、データの本質はほとんど変わっていないのに、そのデータが学習者コーパスと呼ばれるようになると、単語や文のような、言語単位を観察単位にした研究が行われるようになる。その理由は、学習者コーパスもコーパスの一種であるため、コーパス言語学で先行して行われている均衡コーパスの分析法を見習って分析しているためだと思われる。コーパスという名称に惑わされることなく、学習者コーパスでは学習者を観察単位とした分析を行うことが有効だと考えられる。

第５節　本研究と隣接する研究分野との関係

コーパス言語学と隣接する研究分野では、文書や執筆者を観察単位とすることが前提となっている研究が多い。伊藤（2002:3-18）によると、「数学の方法を使って言語を研究する分野全般」を（広義の）数理言語学といい、その下に（1）計量言語学、（2）（狭義の）数理言語学、（3）計算言語学が分類されるという。コーパス言語学は（1）の計量言語学の下に位置づけられている。（1）～（3）の研究内容は以下のとおりである。

（1）計量言語学―計量語彙論，文体統計学，言語年代学，言語行動の計量的調査・研究など
（2）数理言語学―形式文法論，形式意味論
（3）計算言語学―コンピュータ利用の言語研究，自然言語処理の研究（機械翻訳研究）など　（伊藤，2002:4）

この中でコーパス言語学と同様に、計量言語学の下位分類に位置づけられている文体統計学の分野では、文書や執筆者を観察単位とすることが基本となっている。文体統計学のさらに下位分野である文章心理学の研究に安本（1960）がある。ここには、志賀直哉の「暗夜行路」と有島武郎の「或る女」を使用し、句点、直喩、声喩、色彩語の出現の仕方などを比較した研究（pp. 29-62.）や、源氏物語の宇治十帖の著者推定を行った研究（pp.77-137.）などが収載されている。また安本（1965）では、筑摩書房の『現代日本文学全集』を用い、100 人の作家から一編ずつの作品を取り上げて、直喩、声喩、色彩語の頻度とセンテンスの長さ、会話文の量など、15 の特性を使用した文体の比較分析が行われている。これらは文書を観察単位とし、その著者の文体的特性を論じた研究である。

また、同じく文体統計学の一分野である計量文献学の研究に村上（2002）がある。ここでは、井上靖、中島敦、三島由紀夫、谷崎潤一郎などの 4 作家 9 作品を使用して、どの助詞の後に読点を打つかという頻度をもとに作品のクラスター分析を行った研究や、助詞「にて」や接頭語「大」の頻度、形容詞÷接尾語の比率などから、日蓮が書き残したとされる日蓮遺文の真贋鑑定が行われている。これらの観察単位も文書である。

さらに計量文献学で使用される研究手法に焦点を当てたテキストマイニングの概説書に金（2009）がある。テキストマイニングは「文字列で記述されたテキストデータの山から情報や知識を探し出すことを目的とした分野」（金，2009:ｖ）であり、ここでも

基本的に文書が観察単位となっている。

また、近年、フリーの統計ソフトである R[7]を使用したテキストマイニングの概説書が続々と刊行されている（石田・小林，2013；樋口，2014；石田，2017；小林，2017a；小林，2017b など）。これらでは、対応分析、多次元尺度法、ネットワーク分析、クラスター分析、決定木、ランダムフォレストなどの分析手法や、Web 上の SNS からテキストデータを取得するウエブスクレイピング、トピックモデルを使用した文書分類など、計量文献学で扱ってきたテキストマイニング技術だけでなく、新たに自然言語処理の技術を利用して開発された分析法が紹介されている。これらの全ての分析単位が文書であるとはいえないが、その多くは文書を観察単位とした分析法である。

文章心理学や計量文献学は基本的に作品を通して個別の作家の文体的特徴を明らかにする研究分野であり、英語や日本語という巨大な母集団の一般的な特徴の解明を目指すコーパス言語学とは基本的な目的が異なる。しかし、これまでコーパスに集積されているテキストを全体で一つの文書であるかのようにとらえてきた Bag of Words のコーパス観から、コーパスをさまざまな言語タイプの特徴や、書き手や話し手の個性が反映された文書の集合体と捉える本研究のコーパス観にシフトする場合、これらの研究分野で培われてきた分析手法は非常に有益であり、学ぶべきところが多い。

ただし、本研究で扱うのはこれらの手法や最新のテキストマイニングの技術を取り入れた分析法ではなく、それ以前のごく基本的な分析法である。現在、コーパスを使用した言語研究においても最新のテキストマイニング技術の利用が始まりつつあるが、コーパスを使用した言語分析にただテキストマイニングの技術を適用するだけでは、またもや「テキストマイニング技術のブラックボックス的な使用」になりかねない。テキストマイニングの技術を、コーパスを使用した言語研究に有効に適用するためには、これまでのコーパス観から、コーパスをさまざまな言語タイプや執筆者の個性が反映された文書の集合体と捉えるコーパス観へのシフトが必要である。その意味で本研究は、コーパスを使用した言語研究にテキストマイニングの技術を有効に取り入れるための理論的背景を与える研究に位置づけられる。

第 6 節　先行研究の問題点と解決すべき課題

本節ではこれまで指摘してきた先行研究の問題点をまとめ、改めて本研究が解決すべき課題とは何かを確認する。

[7] http://www.r-project.org/

無原則に集められた言語データに比べ、均衡コーパスに信頼性があるのは、均衡コーパスが研究対象となる言語の母集団に対して、代表性を持っているためである。この代表性を確保するための手段が無作為抽出である。このため、均衡コーパスでは、母集団を定め、そこから標本を無作為抽出する方法でコーパスが構築されている。母集団から無作為抽出された個体の観測値を分析することによって母集団の性質を推測することが、統計分析の目的である。

問題は、コーパスの構築に当たって、何が無作為抽出されたのかにある。たとえば、Brown コーパスの母集団は、アメリカで 1961 年に出版された出版物である。Brown コーパスではこのリストから 500 の出版物が無作為抽出され、さらにそこから各 2,000 語の文書が無作為抽出されている。基本的に考えれば無作為抽出されたのは出版物（の一部の文書）である。しかし、文書という言語の塊を集落抽出したと考えると、文書を構成する文字、単語、文などを無作為抽出したと考えることもできる。このため、先行研究では文字、単語、文などの言語単位を個体とする統計分析が行われてきた。

しかし、言語データの場合、集落抽出法の精度は低いことが知られている。また、集落抽出された単語の分布は、単純無作為抽出した場合の二項分布には従わない。この理由は、言語は文法規則に則り、意味的なまとまりを持って使用されているためである。集落抽出法で抽出された言語単位を個体と考えることは、統計分析の前提となっている独立性の仮定に違反している。データに独立性がないと、統計分析の結果に深刻な影響が出る。このため、先行研究のように、文字、単語、文などの言語単位を個体とする統計分析を行っても、必ずしも有効な分析にならないと考えられる。この問題は学習者コーパスを使用した研究でも同様で、学習者コーパスに出現した単語、文、コロケーションなどの言語単位を独立した個体と考えることはできない。これが先行研究における問題点である。

これを解決するには、先行研究の分析法に替わり、文書や学習者を個体とする統計分析を行う必要がある。しかし、コーパスを使用した言語研究において、文書や学習者を観察単位とする分析は、これまで体系的には行われてこなかった。このため具体的にどのような方法で分析を行えば、言語学的にも有効な分析になるのかを検討し、その方法を提案していく必要がある。これが本研究で解決すべき課題である。

第 3 章　本研究が対象とする分析法の概要

本章では、文書を観察単位として分析する場合、どのような手順で分析を行えば有効な分析ができるのか、本研究で扱う分析法を何にするのかという分析法の枠組みについて検討する。第 1 節では、この検討に入る前に、本研究の分析に使用する「頻度」と「文書度数」という用語を定義する。第 2 節では、本研究で扱う分析法とその手順について検討する。

第 1 節　頻度と文書度数の定義

はじめに、分析に使用する用語を定義する。コーパス言語学では、コーパスに出現した文字、単語、文などの言語単位の個数を頻度と呼んでいる。しかし、本研究ではコーパスにおける個体は文書と考え、個体に出現した言語単位の観測値は変数と考える。つまり単語の個数は値であるから、これを頻度と呼ぶことは適切ではない。

> ある値を示したデータの個数（対象の数）を，その値の**度数**（frequency）といいます。たとえば，“3”というデータが全部で 12 個あったら，そのまま，“3 という値の度数は 12 である”といいます。質的変数の場合には，“各カテゴリーに分類されたデータの個数”と表現した方が分かりやすいかもしれません。そして，どのような値（ないし，カテゴリー）のデータが何個ずつあるのかといった，各値とそれぞれの度数を対応させたものを，**度数分布**（frequency distribution）といいます。　（吉田，2001:27）

これに従えば、単語の個数はある文書に出現した「値」であり、たとえば 12 文書に単語 A がそれぞれ 3 語ずつ含まれていた場合、「単語 A の値が 3 の度数は 12 である」という言い方が適切である。しかし、コーパス言語学では言語単位の出現回数を「頻度」と呼ぶ習慣が定着しているため、これを異なる名称で呼ぶとかえって混乱を招く恐れがある。このため、本研究では言語単位の個数はそのまま「頻度」と呼ぶ。その上で、本来「頻度」も「度数」も同じ frequency を意味するものの、ある値を示した文書の個数は「文書度数」と呼び分け、これらに関する用語を以下のように定義する。

(1)　頻度：言語単位の個数。基本的にはコーパスに出現した全個数を指すが、あるジャンルやある文書に出現した個数も頻度と呼ぶ。また、言語単位を単語に限定する

場合は「単語頻度」、文に限定する場合は「文頻度」などと呼ぶ場合がある。

(2) 文書度数：ある値やあるカテゴリーに合致した文書の個数。「単語の値が 3 の度数は 12 である」という言い方の場合、「単語頻度が 3 の文書度数は 12 である」のように表現する。

(3) 文書数：一般的な文書の個数。たとえばコーパスで検索したある単語の合計頻度が 50 である場合、10 文書で 50 になる場合もあれば、20 文書で 50 になる場合もある。この場合の 10 文書や 20 文書はすべてが同じ頻度の文書とは限らないため、文書度数ではなく、文書数と呼ぶ。

第 2 節　本研究で扱う分析法の内容と手順

本節では、本研究で扱う分析法の内容と手順について検討する。統計分析の入門書では、データ分析に際して行うべき最も基礎的な手順として、①データに異常がないかどうかを観察するデータクリーニング、②度数分布表の記述と観察、③代表値を使用した数値要約によるデータの記述と観察をあげているものが多い（Woods, Fletcher, & Hughes, 1986；吉田，2001；市原・岩本，2006；小島，2006；石川・前田・山崎，2010 など）。②、③は記述統計と呼ばれ、観測されたデータの特徴を分かりやすい形で記述・観察する分析法である。文書を観察単位とした分析においても、これらの手順を踏んで分析を行うことが有効だと考えられる。

データを分析する前に行われる作業がデータクリーニングで、たとえば質問紙調査であれば、協力者がまじめに記入していないため、どの項目も同じ番号になっているとか、質問文が適切でなかったため一つだけ選択してほしい設問で選択肢が複数選択されている項目があるなど、何らかの異常を見つけて対処することをいう。コーパスでも、データに全く異常がないことは考えられないため、データクリーニングを行う必要があるが、大規模に集積されたコーパスの場合、使用者がそのデータの全てに目を通してチェックすることは実質的には不可能である。また、全ての文書を読んでチェックしたとしても、どのような文書を異常だと判断するのか、認定基準を作ることも難しい。

データクリーニングの対象を検索した用例だけに絞っても、その用例が大量である場合、それを目視で丹念にチェックすることは困難である。このため、小規模なコーパスで文字列検索を行った場合などでは、調査対象外用例のチェック等のデータクリーニングが行われているが、BCCWJ のように形態素解析済みの大規模コーパスの場合、それほど丹念なデータクリーニングは行われていないのが実情だと思われる。また、たとえ

目視による用例チェックを行ったとしても、検索された用例を文単位で読むだけでは、用例を文書単位でまとめた場合にはじめて明らかになる異常を発見することは困難である。

しかし、文書を観察単位とした分析法では、外れ値（「データ集合の中で他の観測値と全く異なる観測値」，Upton & Cook，2011:301）となっている文書に異常が見られることが多いため、文書の度数分布を観察することで、異常な個体を発見しやすい。外れ値となっている文書が見つかった場合、その文書の内容をチェックすることによって、異常なデータなのか、特異ではあっても正常なデータなのかを判断することが可能になる。つまり、コーパスのデータの全てに目を通すのは不可能でも、分布観察で外れ値となっている文書が特定できれば、効率よくデータチェックを行うことができる。一般的な統計分析の手順では、データチェックを行った後に度数分布の観察へと進むことが多いが、コーパス分析の場合、文書度数分布を観察しながら、データチェックを同時に行う方法が有効だと考えられる。

この文書度数分布の観察では、外れ値の発見だけでなく、当然ながら調査対象の言語現象がコーパス内にどのように分布しているかという特徴が観察できる。また、特徴的な文書のテキストを選抜して分析することで、そのような分布の特徴がなぜ生じるかについても、ある程度の予測が可能になる。従来の言語単位を観察単位とした分析法では、調査対象の頻度は十分なチェックを経ないまま、母集団を反映した頻度という信頼性が仮定されて分析されるが、文書を観察単位とした分析法では、その頻度に異常がないかどうかをチェックし、なぜそのような頻度になるのかという理由についても、ある程度把握しながら分析することが可能になると思われる。

しかし、コーパスから何らかの対象を検索した場合、出力されるのは用例単位のデータである。まず、これを文書単位に集約する必要がある。また、文書の語数が均一でない場合、粗頻度で度数分布をまとめても有効な分析はできない。語数の違いを考慮に入れながら、分布を観察する必要がある。これらの方法は先行研究で十分に明らかにされているとはいえないため、分析の第一段階として、どのような分析法を用いれば有効な度数分布観察ができるかを、明らかにしていく必要がある。

度数分布観察の次の段階は、代表値によるデータの特徴記述である。これは、平均値や中央値などの代表値によってデータを要約し、特徴をつかみやすくする分析である。文書を観察単位とした場合、たとえば文書 A には調査対象が 3、文書 B には調査対象が 5 出現するなど、文書ごとに調査対象の観測値（変数）が異なる。このため、平均値や

ばらつきの指標である標準偏差などが算出できる。しかし言語単位を観察単位とした場合、個体ごとに変化する観測値は存在しない。あえていうなら個体の値は検索によって出現したという意味で、全て 1 である。この場合、個体の平均値も中央値も 1 になるため、平均値や中央値を使用する分析は考えにくかった。これに替わって行われてきたのが、カテゴリーごとに算出した調整頻度による比較である。これはカテゴリーごとに個体数を合計し、カテゴリー間の比較が可能になるように語数を調整した値で、つまるところカテゴリーの合計数の比較である。しかし文書を観察単位とした分析法では、個体の観測値（変数）はそれぞれ異なるため、そのばらつきや分布を考慮に入れた平均値や中央値の分析も可能になる。このため、このような代表値による分析の有効性を検討してみることが必要である。

これらの分析によって言語学的に意味のある頻度差が見つかった場合、その次に求められるのは、その頻度差がそのコーパスだけに限定される違いなのか、それとも母集団でも成立している違いなのかを推測する統計的検定であろう。コーパス言語学では、カイ二乗検定がこれまで最も多く使用されてきた。しかし、従来の単語や文などの言語単位を観察単位とした分析法では、独立性の仮定を満たしていないため、有効な検定ができないと考えられる。このため、文書を観察単位とした場合、従来のカイ二乗検定の問題が克服できるかどうかを検討し、コーパス分析に有効なカイ二乗検定の方法を明らかにする必要がある。

言語学的に意味のある頻度差が母集団でも成立する可能性が高いと判断できた場合、その次に求められるのは、そのような頻度差がなぜ生じているのかという因果関係の解明だと思われる。相関関係に基づいて行われる回帰分析は、このような因果関係の解明で最も多用されている分析法である。このため、回帰分析においても、従来の分析法のように言語単位を観察単位とするのがよいか、本研究のように文書を観察単位とするのがよいかという問題を検討し、コーパス分析で有効な回帰分析の方法を明らかにする必要がある。

以上のように、度数分布の観察から回帰分析までの方法が明らかになれば、文書を観察単位とした分析法の基礎ができあがる。このため、本研究においては、「文書分布と代表値の観察による有効な頻度差の発見（第 4、5 章）→それが母集団でも成立する可能性の検定（第 6 章）→頻度差が生じる因果関係の解明（第 7 章）」という枠組みで研究を行うこととする。これらの分析では、それぞれ言語単位を観察単位とした従来の分析法と本研究が主張する文書を観察単位とした分析法のどちらが有効であるかについ

ての考察を行い、コーパスを使用した有効な言語分析とはどのような方法であるかを明らかにしていく。

第 4 章　分布観察の方法

本章では、調査対象の文書分布を観察することにより、データクリーニングが必要な外れ値を発見するとともに、調査対象の言語的特徴を把握する方法を検討する。データには BCCWJ 図書館 SC を使用する。第 1 節では文書の文字数がほぼ 1,000 字に固定されている固定長を使用し、文書度数折れ線を描くことで分布観察を行う方法を検討する。第 2 節では、文字数が文書ごとに異なる統合形式を使用し、語数と頻度の散布図を描くことで分布観察を行う方法を検討する。固定長の文書度数分布図と統合形式の散布図では、同じ単語でも分布の状態に違いが見られる。そこで第 3 節では固定長と統合形式の文書内で単語がどのように出現しているのかを観察し、固定長と統合形式のどちらが母集団の縮図に近い分布となっているかについて考察する。第 4 節では文書ごとの語数を一定に平準化する個別調整頻度について考察し、統合形式を使用して文書度数折れ線を描く方法を検討する。第 5 節では、そもそもどれぐらいの文書数があれば母平均の正確な推定が可能になるのかという必要文書数の見積もりを行い、コーパスを使用して有意義な分析ができる調査対象の目安について考察する。最後に第 6 節で本章のまとめを述べる。

第 1 節　文書度数折れ線による固定長の文書分布観察

本節では、文字数がほぼ一定である固定長を使用し、文書度数折れ線を描くことによって文書分布を観察する方法を検討する。

文書度数分布図とは、すべての文書における調査対象の頻度（観測値）に対して、単語頻度 0 の文書度数、頻度 1 の文書度数、頻度 2 の文書度数、・・・、のように、単語の頻度ごとに文書の度数をまとめて並べたグラフのことである。コーパスで単語を検索すると、その単語の用例が出力される。それらの用例を、その用例が出現した文書ごとに集約すると、文書ごとの頻度が確定できる。度数分布表とは、頻度順にその頻度を持つ文書数をまとめた表であり、それをグラフに描いたものが分布図である。度数分布図はヒストグラムで描かれることが多いが、ヒストグラムの場合、度数が少ないと棒線の存在が分かりにくいため、本研究では度数分布を度数折れ線（折れ線グラフ）で図示する。

度数分布図の観察は、調査対象の分布の特徴を捉えるために行われる。コーパスで無作為抽出された個体を単語と見なすと、単語はコーパスという巨大な単語の袋（Bag of Words）にランダムに散らばって存在していると考えられる。これがコーパスにおける

最も素朴な単語分布のイメージである。しかし、実際のコーパスは文書の集合体であるため、単語が分布しているのではなく、文書が分布している。文書には小説や論説文などの種類や児童書、専門書のようにそれが書かれた対象、目的、難易度など、さまざまな特徴が備わっている。このため、調査対象がそれぞれの文書に何語含まれているかという観測値に基づいて分布図を描くと、特徴の似た文書が近くにまとまり、一定の分布を示すようになる。この分布を観察することで、調査対象の特徴を捉えることができる。

第 1 章で引用した小島（2006:17）で、「分布」観察の意義をもう一度確認しておこう。

> （筆者注：図表 1-1 女子大生 80 人の身長（cm）は）「日本人の成人女性」の一部という集団を扱っていますが、属するメンバーの身長は、さまざまな数値をとります。この「**さまざまな数値をとる**」ということを、専門の言葉で「**分布する**」といいます。分布が生じるのは、その数値が決まる背後に何らかの「**不確実性**」が働いているからに、ほかなりません。不確実性のメカニズムが、まちまちな身長の数値を生み出すと考えるのです。ところが、「不確実」と一口にいっても、それらには固有の「特徴」や「癖」があることがわかっています。その固有の特徴や癖を「**分布の特性**」と呼びます。
>
> さて、この身長のデータに固有の特徴や癖は何でしょうか。データ解析になじみのある方なら、数値をじっとにらんでいるだけで、多くの特徴や癖を引き出せるのでしょうが、普通の人にはただの数字の羅列にしか見えないに違いありません。
>
> そこで、この生データ、つまり「生の現実」から、何かその分布の特徴や癖を引き出すための手法が必要になります。それが「**統計**」という手法なのです。
>
> （小島，2006:17）

小島（2006:17-8）では、この統計手法として、①グラフ化して特徴を捉えることと、②一つの数字で特徴を代表させることをあげている。これからも分かるとおり、度数分布表を作図して観察することは、その図から分布の特徴を捉えるために行われる。

この度数分布図を観察することで、どんな特徴が明らかにできるのか、以下で具体的な調査対象の分布を観察してみる。調査対象は、選択が恣意的にならないように、BCCWJ 図書館 SC・固定長・長単位の語彙頻度表で順位 1 位の格助詞「の」、100 位の代名詞「そこ」、500 位の名詞「社会」、998 位の動詞「戦う」、3,010 位の形容詞「興味

深い」、9,890 位の名詞「乗り物」、20,000 位の名詞「虫歯」を取り上げる。順位が半端な単語があるのは、同順位の単語が複数あって、1,000 位や 3,000 位など、区切りの良い順位が存在しないことによる。

図 4.1 は、頻度順位 1 位の「の」の度数折れ線である。「の」の場合、図書館 SC10,551 文書のうち、1 回も出現しない文書は一つだけで、残りの文書にはすべて出現する。頻度合計は 340,436、1 文書当たりの平均は 32.3、標準偏差 10.3 である。標準偏差を平均値で割った値で、ばらつきの大きさを表す変動係数は 0.32 と、非常に小さい。つまり、文書ごとの観測値はばらつきが小さく、概ね似た値になっている。固定長の目安となっている 1,000 字は、文庫本にすると約 2 ページ強の長さである。図書館 SC の長単位平均では約 523.6 語になる。図書館に収蔵されている書籍では、文庫本約 2 ページ当たりで「の」を 32 回程度使用しているものが多く、そこから±10 回程度の範囲で全ての書籍の 70%近くを占めると考えられる。

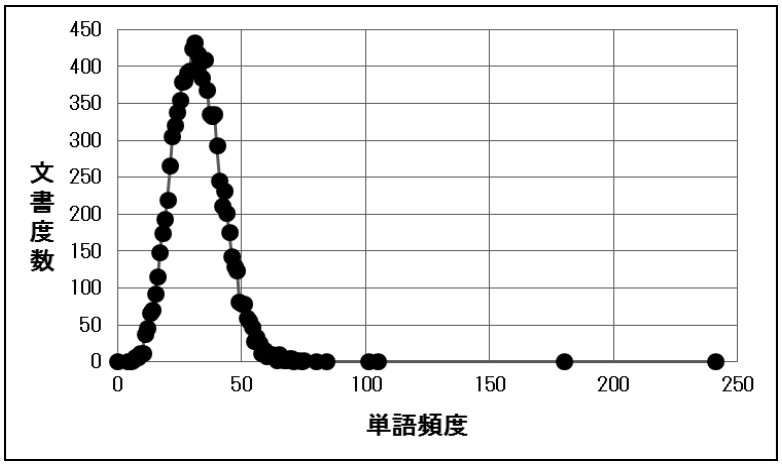


図 4.1　「の」の文書度数折れ線

図 4.1 から分かることは、日本語の書き言葉において「の」の頻度はかなり似通った使われ方をしているということである。その一方で、中にはほとんど「の」を使用しなかったり、平均の 2 倍以上も使用したりする文書も存在し、「の」の使用には多様性がある。分布図は正規分布に近い形をしているが、右裾が厚く、平均より「の」を多用する文書の方が多い。「の」の頻度は、このように、どのような頻度の文書がどのように分布しているかによって決定される。コーパスを Bag of Words のように考えると、「の」は単語の袋の中にまんべんなく散らばっているように思われるが、実は、図書館 SC の中にどんな言葉遣いをする文書がどのように分布しているかによって決まる値なので

ある。

次に、図 4.1 における外れ値の検討を行う。図 4.1 では、「の」が出現しない文書が 1 文書存在していた。(1)は、この文書の一部である。

(1) 拝啓先年御世話被下候千住ノ碑出來候由安心仕候黒木君へハ當時小生ヨリ薄儀呈置候處千住ヨリハ「御心安キ中ユヱ謝儀ハ不仕」ト申來候千住ヨリ來次第黒木君へモ改テ御禮可申考居候ニ右ノ如クニシテ少シアテガハヅレ候

（LBd9_00189, 森鷗外,『鴎外全集』）

この文書は、森鷗外の書簡で、漢字カタカナ交じり文で書かれている。BCCWJ では漢字カタカナ交じり文がうまく形態素解析できておらず、単語の多くが「カタカナ文」という品詞名で解析されている。(1)の 1 文目を例にすると、「拝啓先年」「御世話被下候」「千住」「ノ」「碑出來候」のまとまりで長単位として認定され、その品詞名がカタカナ文になっている。これは、形態素解析の誤りだと思われる。このような文書では研究者の判断で「の」の頻度を数え直すか、この文書は外れ値と考えて分析から除くことを検討する必要がある（本章では分布観察が目的であるため、第 2 節以降の分析でも、外れ値を含めて分析を行う）。

BCCWJ には、この他にも「の」がカタカナ文として解析されている単語を含む文書が 36 文書あり、文書の全体が漢字カタカナ交じり文で書かれている文書から、文書の一部に漢字カタカナ交じり文が出現するものまでさまざまなパターンが存在している。ちなみに「の」の頻度が少ない 2 番目の文書も漢字カタカナ交じり文であり、これらを除いて最も「の」の出現が少ない文書は、山田太一『ふぞろいの林檎たち』（LBf9_00007）の 5 回である。この文書は、会話を主体としたテレビドラマの脚本であることから、「の」の出現が少なくなっていると考えられる。

頻度が少ない文書とは反対に、図 4.1 では頻度が 180 と 241 と、非常に多く出現している文書も観察される。頻度 180 の文書は、吉増剛造『この時代の縁で』（LBm9_00253）で、固定長・長単位の語数が 2,098 語である。頻度 214 の文書は、ジェイムズ・ジョイス（著）丸谷才一（訳）『ユリシーズ』（LBr9_00057）で、語数が 4,108 語である。固定長では文が途中で切れるのを避けるため、1,000 字を超えて文が終了するところまでが抽出されている。この二つの文書は、どこが文末か簡単には判断できなかったため、長大な長さになってしまった固定長である（田野村 2014:124-5）。次の用例(2)、(3)はこの

文書の一部である。

(2) “…仕草”“子供の挙動”は、柳田民俗学の生命／芯のようなところ、…そうか『山の人生』冒頭の“お父、これでわたしたちを、…”“あの斧”をとぐ“仕草”“挙動”の“芯のようなところ、…”―おそらくこの“芯”はわたしの誤用といってもよい“語の濫用”で絲玉の芯、蹴鞠の芯、それをほどいて行くときの心の弾みまでも込めようとしている。

（LBm9_00253，吉増剛造,『この時代の縁で』）

(3) それにああいう所のにおいと来たらカマフォード家のパーティからポールディといっしょにかえって来た晩オレンジとレモネードのせいでおしっこが出たくなってこまってああいうのの１つにはいったひどくさむい夜だったのでとてもこらえきれなくてあれはいつだったかしら九十三年ね運河がこおっていたっけｙｅｓ

（LBr9_00057，ジェイムズ・ジョイス（著）丸谷才一（訳),『ユリシーズ』）

これらの文章を読むと、特に(3)では句読点が使用されていないため、外形からは文末がどこか分かりにくいが、意味的に文末がどこかを判断できないわけではない。固定長では統計分析がしやすいように約 1,000 字で区切るという原則を破ってまで、これらの文書を長く抽出しなければならない理由は、この用例からは考えにくい。このような文書は、外れ値と考えて分析から除くか、第 4 節で述べる個別調整頻度に平準化して分析することを検討する必要がある。

これ以外にも、頻度 105 の文書は、語数が 705 語のギィ・スカルペッタ（著）本多文彦（訳）『小説の黄金時代』（LBr9_00276）、頻度 101 の文書は、語数が 900 語の梶村太市『相続の法律相談』（LBo3_00088）で、図書館 SC 固定長平均の 523.6 語からすると、かなり語数が多めの文書である。

図 4.2 は図書館 SC・固定長・長単位 10,551 文書の語数を、少ない文書から多い文書の順に表示させたグラフである。文書の大半は 500 語を少し超えた語数だが、これより明確に語数が少ない文書と語数が多い文書が交じっている。特に語数が多い二つの文書が先に示した用例(2)、(3)の文書である。このように平均から明確に語数が異なる文書は分析から除くか、これらを含める場合は固定長ではあっても粗頻度ではなく、個別調

整頻度を使用するなどの対応を検討する必要があるだろう。

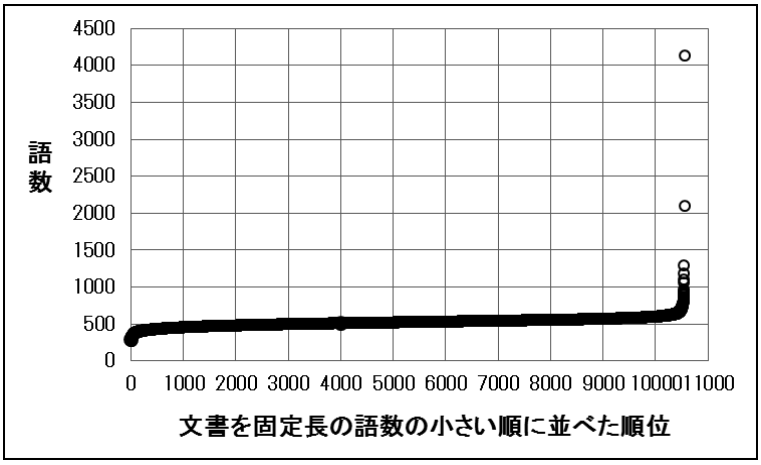


図 4.2 図書館 SC 文書の固定長語数

以上で、「の」の文書分布の観察を終えるが、これに続いて、頻度順位 100 位の代名詞「そこ」の観察に移る前に、度数折れ線の観察によって外れ値が見つかる典型的な例をもう一つ紹介しておく。次の図 4.3 は、図書館 SC・固定長・短単位で「色素」という単語を調査した文書度数折れ線である。短単位で調査しているのは、長単位の場合、「〇〇色素」のように「〇〇」がついた複合語になると、これ 1 語で別の単語として扱われるため、ここで取り上げるような異常が見つけにくくなるためである。なお、図 4.3 では、頻度 0 の文書数が極端に多く、これを含めて文書度数を描くと見にくくなるため、これを除いて図示している（これ以後の度数折れ線も同様である）。

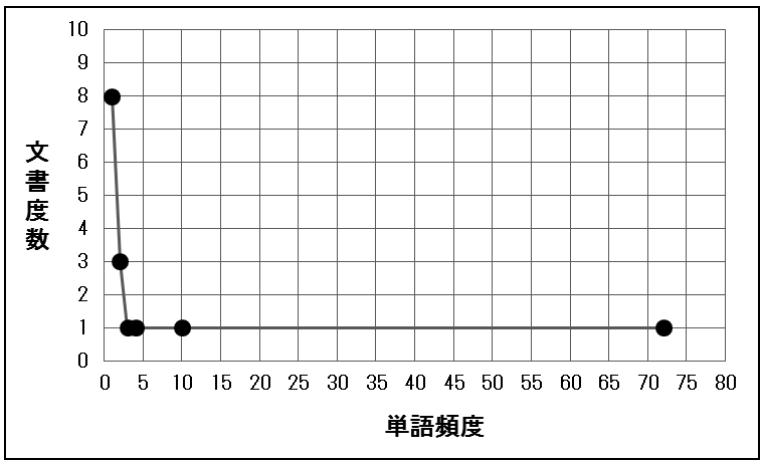


図 4.3 「色素」の文書度数折れ線

図 4.3 では、一つの文書だけ頻度 72 と極端に頻度が高い外れ値が観察できる。図 4.3

の文書数は 15、頻度合計は 103 である。外れ値が存在するため、平均 0.010、標準偏差 0.711、変動係数 72.79 となり、ばらつきが非常に大きい。外れ値の 1 文書で全体の頻度の 7 割を超える。このように少数の文書だけが他の文書とは極端に異なる傾向性を持って出現している場合は、何か異常なことが生じている可能性が高い。

次の(4)はこの文書の用例の一部である。

(4) 従来とは異なった新しい製造方法で生産された天然着色料は，起源，製法が異なるので，使用に際しては新たに申請する必要がある．表　４－十六　化学構造により大別した天然着色料（１）　カロテノイド系アナトー色素，エビ色素，オキアミ色素，オレンジ色素，カニ色素，クチナシ色素，コーン色素，サフラン色素，抽出カロテン，トマト色素，パプリカ色素，ファフィア色素，ヘマトコッカス藻色素，マリーゴールド色素（２）　アントシアニン系赤キャベツ色素，赤米色素，シソ色素，チェリー色素，ダークスイートチエリー色素，モレロチェリー色素，ハイビスカス色素，ブドウ色素，プラム色素，紫イモ色素，紫コーン色素，紫ヤマイモ色素，ウグイスカグラ色素，（以下略）

（LBg4_00002，谷村顕雄，『食品添加物の実際知識』）

(4)は最初の 1 文目の後に書かれている「表　４－十六　化学構造により大別した天然着色料」がキャプションで、これ以降が表の中身である。（１）は表の項目分類で、引用は（２）までで留めたが、固定長の文書では項目数が（９）まであり、まだまださまざまな色素名の列挙が続いている。このためこの文書における「色素」の頻度が 72 という高さになっているのである。

このような表の中身がコーパスのデータとして抽出されているのは、当然のことながら BCCWJ の設計方針でそのように規定されているからである。書籍や雑誌、新聞などの印刷物からテキストを抽出する作業はさまざまな困難を伴う。中でも印刷された紙面の中で、何が抽出するべきテキストなのかを判定することは意外に難しい。紙面構成が単純な小説などの場合は、そのまま一続きのテキストを抽出すればよいが、紙面に写真や図、表などが含まれていた場合、日本語とはいえない写真や図を除外するのは当然としても、その写真等につけられているキャプションは抽出するのか、表の中に記述されているテキストはどうするのかといった問題が生じる。この基準については、柏野・丸山・稲益・田中ほか（2009）に詳しく記述されているが、特に問題となるのが表の中の

テキストの抽出方法である。次の引用は、これをどのような考え方に基づいて判断していくかの原則が記されている記述である。

> 第 I 部 3.3.2 節でも述べたとおり，図 4-24 に示すような「行列見出しを備えた表」を「表」の典型と考える。さらに言えば，より典型的な表は罫線が引かれている。このような表は先述のとおり，文字列を一方向に読むことができない。そのことを根拠に文字列が図式化されている「フィギュア」の類型の一つと考える。
>
> しかしながら，第 I 部 3.3.2 節に，「一見フィギュア本体に見える要素であっても，その内部にある言語表現を一方向に読み進めることができれば，フィギュア本体とは見なさず，排除の対象とはしない」ことに注意が必要であり，その根本には「印刷紙面上に現れた文字列は，それが現代日本語として読み進められる限り，できるだけサンプルとして収録する」という姿勢があると述べた。このことがもっともよく問題になるのは，「表」の認定においてである。
>
> サンプリングする紙面には，「表」，あるいは「表のようなもの」が数多く出現する。それらのうち，一方向に読むことが十分可能である文字列が，ただ罫線で囲んであるだけで，なおかつ，「図表」などと明記されている場合がある。しかし，それらは「図表」とは認定せず，積極的にサンプリング対象要素と指示すべきものであると考える。逆に，「図表」という明記はなく，場合によっては本文中に入り込んでいるようなものでも，一方向に読み進めがたい，図式化された文字列は，「表」と認定し，積極的に排除要素と指定すべきものであるとも考える。
>
> サンプリングする紙面には，典型的な「行列見出しを備えた表」ではない「表のようなもの」が数多く出現し，その判断はしばしば難しい。
>
> （柏野・丸山・稲益・田中ほか，2009:79）

引用にある「図 4-24 に示すような「行列見出しを備えた表」」とは、4 行 3 列のクロス表で、表内は数値ではなく文字列になっている。「フィギュア」とは、「本文中に含まれている写真や図など，言語表現以外の内容が主たる対象となっている部分」（柏野・丸山・稲益・田中ほか，2009:21）を指す。上記を見ると、典型的なクロス表のように、表組自体に意図があるものはフィギュアと認め、表の形はしていても、一方向に読み進

めることができればフィギュアではないと認定する。このため、(4) の「表　４ - 十六化学構造により大別した天然着色料」は、一方向に読み進めることができるという理由で、サンプルとして抽出されたと考えられる。

しかし、この文書を抽出した『食品添加物の実際知識』という書籍の全体を考えた場合、この表を含めた方がこの書籍の縮図となっているのか、含めない方がより縮図に近いのかといえば、表を含めない方が全体の縮図に近かったと思われる。この表に記載されている色素名が、書籍全体に渡って何度も出現するとは考えにくい。つまり、書籍の別の部分から固定長を抽出していたら、「色素」の頻度は全く異なったものになったと考えられる。

以上のように、特殊な事情によって頻度が偏っている文書は、文書度数折れ線を観察することで発見が可能である。また、特殊な文書が見つかった場合、この分析法であれば、文書の中身を読んで異常の有無を確かめることができる。このような対応は、従来の言語単位を個体と考えてきた分析法では難しかった対処法である。

さて、ここで長単位の頻度順位に基づいて選択した単語の文書分布の観察に戻る。次の図 4.4 は、100 位の「そこ」、図 4.5 は 500 位の「社会」の文書度数折れ線である。「そこ」の頻度合計は、4,289 である。この頻度は一見多く思えるが、出現する文書数は 3,143（全 10,551 文書中 29.8%）で、1 文書当たりの平均は 0.41、標準偏差 0.73、変動係数 1.78 となっている。「そこ」は約 1 万冊の書籍の文書のうち 3 割程度にしか出現せず、ばらつきも大きい。

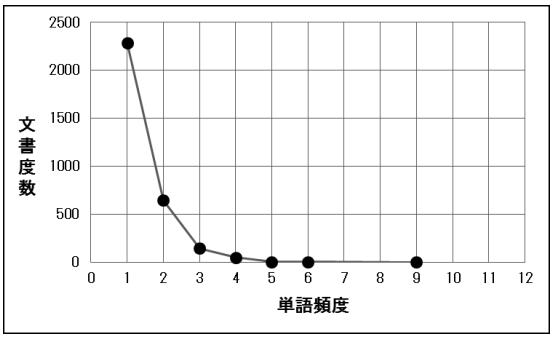


図 4.4　「そこ」の文書度数折れ線

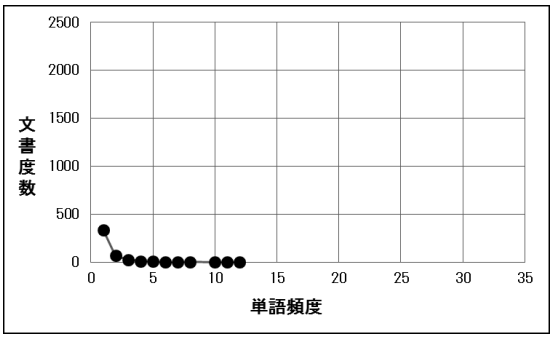


図 4.5　「社会」の文書度数折れ線

「そこ」のようにごく基本的な単語であれば、母集団ではおそらくほとんどの書籍に出現し、その分布も正規分布に近似すると思われるが、1 冊の書籍から抽出している文書の範囲が約 1,000 字という短い区間であるため、「そこ」が出現する文書は全体の 3

割程度になっている。

これが「社会」になると、出現する文書は非常に少なくなり、頻度合計 762、文書数 472（全 10,551 文書中 4.47%）、平均 0.072、標準偏差 0.446、変動係数 6.19 となっている。図書館 SC・固定長・長単位の異なり語数は 219,124 語である。「社会」という単語は、およそ 22 万語ある単語のうち、順位 500 位に入るほどの高頻度語だが、ほとんどの文書には出現していない。そもそも「社会」という話題語の場合、機能語の「の」や「そこ」とは異なり、母集団の書籍全体で検索しても、この単語が 1 回も出現しない書籍も多いと考えられる。

つまり、あらゆる文書に出現するような超高頻度の機能語の場合、文書度数分布は概ね正規分布に従うと考えられるのに対し、話題語の場合、母集団の図書館 SC 全体で調べても文書度数折れ線は図 4.5 のような L 字型の曲線になっていると考えられる。母集団の図書館 SC に「社会」という単語がどれぐらい出現するかは、日本人の文化における「社会」という単語の重要度が絡み合った中で決定される値で、もはや言語の問題というよりは、日本文化の中に占める単語の地位を表しているとも考えられる。

図 4.5 の場合、頻度が 1 の文書であれば、たまたま「社会」という単語が出現した文書である確率が高いであろう。これが頻度 10、11、12 などの文書なら、「社会」に関するテーマで本格的に書かれた書籍の一部に出現している可能性が高いと思われる。(5) は、頻度 12 の文書の一部である。

(5) マルクスは十九世紀の産業革命の最中のイギリスに不正義と不平等の社会を見て、そういう社会を転覆せしめて正義と平等の理想社会を作ろうとしたわけであるが、その理想はあまりに非現実的であり、究極的には失敗であったとしても、資本主義社会をそのまま放置すれば欲望肥大の不正義と不平等の社会となるという指摘は、今の日本の現実を見ても決して間違っていない。
（LBj2_00063，梅原猛，『自然と人生』）

「社会」の頻度は、このように「社会」という単語が一切出現しない文書、たまたま出現した文書、本格的に「社会」に関して記述された文書などの組み合わせによって決定される値である。

次の図 4.6 は、頻度順位 998 位の動詞「戦う」の分布である。これぐらいの順位になると同順位の単語が複数あるため、ちょうど 1,000 位の順位が存在しない。998 位が 1,000

位に最も近い順位である。「戦う」は、最大の頻度でも 4 で、「社会」のように頻度が 10 を超える文書は存在しない。

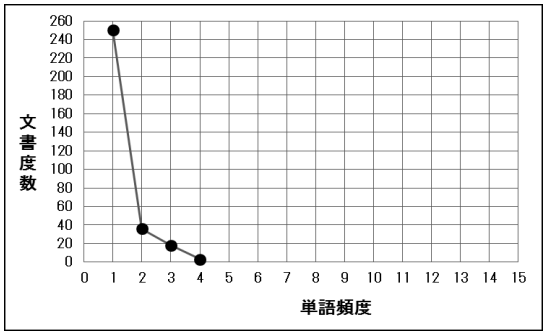


図 4.6 「戦う」の文書度数折れ線

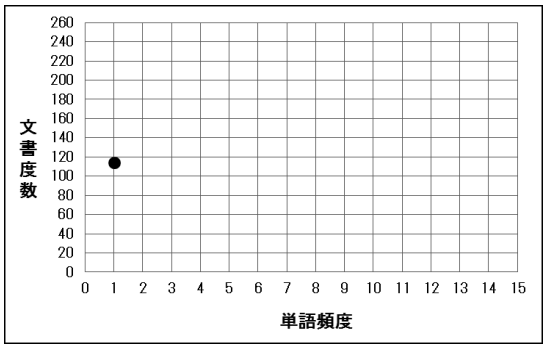


図 4.7「興味深い」の文書度数折れ線

図 4.7 は 3,010 位の「興味深い」で、頻度が 1 の文書しか存在せず、頻度合計も文書数も 114 である。この分布図の結果からすると、文庫本 2 ページ程度の長さの文章で、何度も「興味深い」を連発する文書はほとんどないと考えられる（これについては第 2、3 節で再考する）。

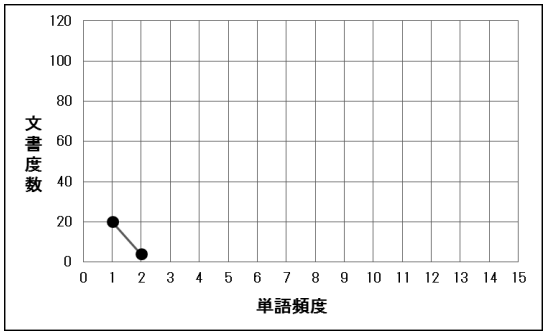


図 4.8 「乗り物」の文書度数折れ線

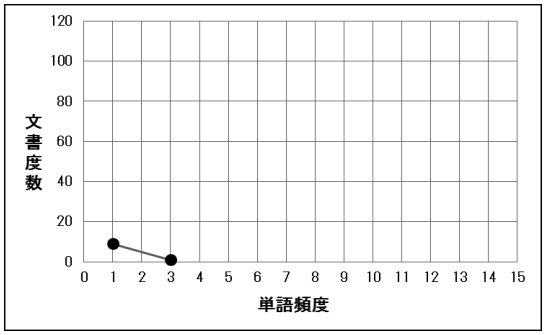


図 4.9「虫歯」の文書度数折れ線

図 4.8 は 9,890 位の「乗り物」で、頻度合計 28、文書数 24、図 4.9 は 20,050 位の「虫歯」で、頻度合計 12、文書数 10 である。頻度順位が下がると出現する文書も頻度合計も非常に少なくなる。「乗り物」も「虫歯」も、「社会」と同じ話題語である。「乗り物」や「虫歯」を話題とした文書であれば、「社会」と同じように文庫本 2 ページ強の範囲で 5 回や 10 回は出てきてもおかしくないようにも思えるが、そのような傾向は見られない。(6)は、「虫歯」が 3 回出現する文書の一部である。

(6)　たしかに日本人の乳歯の虫歯は減りました。しかし、おしゃぶりやガムが消

えたために赤ちゃんから幼児期、そして小学生のころまでにやらなければならない「唾液が出る→唾液を飲む」という舌の筋肉をつけるための運動が行われなくなってしまったのです。

（LBn4_00026，各務肇，『心身の健康をつくる歯の矯正』）

ここでは、「Q７　顎が小さいと歯並びが悪くなるとか。顎の小さいのはやわらかい食べ物のせいでしょうか。かたいものを食べさせて、甘いガムなどは与えないほうがよいのですか」（同じく LBn4_00026 より引用）という質問に答える際に虫歯に言及されており、「虫歯」が話題の中心ではない。世の中には「虫歯」をテーマとした書籍も存在するとは思われるが、そのような書籍が無作為抽出される確率はごく低いためか、「社会」とは分布が大きく異なっている。

以上、文書度数分布の観察を行った。度数分布折れ線を観察することで、外れ値の発見や分布の特徴が捉えられる。この分析法で特に有効な点は、特徴のある文書を特定し、そのテキストを確認できるところにある。文書の中身を読めば、その文書が何らかの異常で外れ値となっているのかどうかが評価できる。また、「社会」と「虫歯」の出現の仕方などについて、高頻度語と低頻度語ではどのように異なるのかの示唆を得ることができる。このように文書度数分布の観察を手掛かりにしてコーパスの文書の中身に分け入り、その特徴を直に確かめることで、コーパスのブラックボックス的使用を避けることが可能になる。コーパス内の全ての文書を読むことは困難だが、特徴的な文書を読めば、ある程度までそのコーパスの資料的性格や、調査対象の言語的特徴を把握することが可能である。

第２節　散布図による統合形式の文書分布観察

本節では文書ごとの語数が異なる統合形式のデータを使用し、語数と頻度の散布図を描くことで文書の分布を観察する方法を検討する。統合形式のデータを使用して、前節と同様の文書度数折れ線を描くためには、第 4 節で考察する個別調整頻度を算出する必要がある。しかし、個別調整頻度は一定語数の中に調査対象が何語出現するかという割合の情報に変換した値であるため、生のデータが持っていた頻度の情報を失ってしまう。また、詳細は第 4 節で述べるが、語数の少ない文書の個別調整頻度を算出する段階で、頻度が過大に調整されてしまう場合がある。このため、文書度数折れ線による分布観察を行う前に、何も加工を施さない粗頻度を使用して、文書の分布を観察しておく必要が

ある。これには、文書の語数と単語頻度の散布図の観察が有効だと思われる。

この散布図を観察することで、どんな特徴が明らかにできるのか、以下で具体的な調査対象の分布を観察してみる。調査対象は、前節でも取り上げた BCCWJ 図書館 SC・固定長・長単位の語彙頻度表で順位 1 位の「の」、100 位の「そこ」、500 位の「社会」、3,010 位の「興味深い」、9,890 位の「乗り物」、20,000 位の「虫歯」とする。

図 4.10 は、図書館 SC・統合形式・長単位における各文書の語数とその文書に出現した「の」の頻度を使用して描いた散布図である。図中の×は、後で言及する文書である。

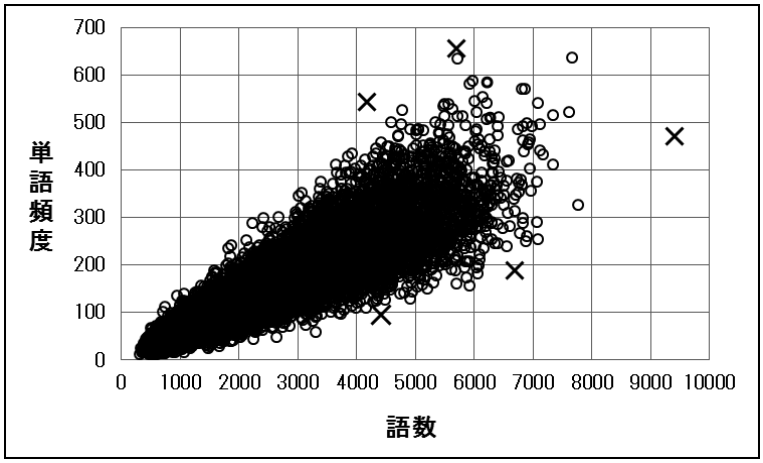


図 4.10　語数と「の」の頻度の散布図（×印は本文で言及する文書）

「の」の頻度には、語数と強い相関があり、ピアソンの積率相関係数は.890 である。ただし、語数が多くなるにつれ、ややばらつきが大きくなる傾向が見える。語数が最も多い文書は右端の×印で、先に用例(3)で観察した『ユリシーズ』である。語数が多い分、「の」の頻度も高くなっているが、語数と「の」の関係については特に異常は見られない。

これに対し、図の中央上部の二つの文書は、語数に対して「の」の頻度が高く、他の文書とは傾向を異にしている。最も「の」が多い用例が(7)、その左側の×印が(8)の用例である。

(7)　十二月、明石の姫君は紫の上の養女として二条院に迎えられ、袴着の儀式が盛大に催される。翌年になって藤壺入道の宮（中宮）は、三月から重く病み、三十七歳を一期として世を去った。

（LBl9_00038，杉山英昭，『「源氏物語」がわかる』）

(8) 〔資料１〕　オゾン層を破壊する物質に関するモントリオール議定書　この議定書の締約国は、オゾン層の保護のためのウィーン条約の締約国として、同条約に基づく、オゾン層を変化させ又は変化させるおそれのある人の活動の結果として生じ又は生ずるおそれのある悪影響から人の健康及び環境を保護するために適当な措置をとる義務があることに留意し、（以下略）

（LBe4_00009，実著者不明，『フロン』）

(7)は、源氏物語を解説した書籍である。文体に特異な特徴はないが、「明石の姫君」、「紫の上」、「藤壺入道の宮」などの登場人物の名前に「の」が使用されている。長単位ではこのような個人名は本来、1単語の固有名詞として単語認定されるのが望ましいが、古典に出現する人名であるためか、人名としては形態素解析されていない。登場人物の名前は、話題語として何度も文書に出現するため、この文書では他の文書には類を見ないほど「の」が多い文書となっている。

(8)は『フロン』という書籍の資料として掲載されている「オゾン層を破壊する物質に関するモントリオール議定書」が、文書として抽出されている。文章を読むと法律の文章に特有の表現で、この引用の後にも「～し、～し、～し、・・・」と、連用中止法が連続して使用され、長く文章が続いている。(8)では、法律文特有の特殊な文体で書かれているため、「の」が多くなっていると考えられる。

一方、「の」が少ない文書には、(9)、(10)のような例が見られる（図 4.10 で(9)は右下、(10)は中央下）。

(9) 横尾―　僕は自意識なんて、屁のツッパリにもならないと思う。細野―　屁のツッパリね（笑）。横尾―　自意識ほど、自分を狭くして、生きにくくしてるものはないんじゃないかしら。細野―　でも、みんなそれを大事にしてますよ、相変わらず。　（LBp7_00030，横尾忠則・細野晴臣，『芸術ウソつかない』）

(10) 姑　何頑張んの？嫁　まあ、とにかく、嫌われないようにっていうか。難しいですけど、言葉で言うのは。なんて言うんだろ？姑　だから、つかず離れずが一番いいのよ。

（LBp3_00160，野村有紀子・野村沙知代，『日本一勇気ある嫁』）

(9)は、横尾忠則と細野晴臣の対談である。(10)も、野村沙知代とその嫁である野村有

紀子のやり取りが書かれていて、会話を書き起こした文章である。両者とも会話文であるため、「の」の頻度が少なくなったと考えられる。

これらの文書を外れ値と考えるか、特異ではあっても外れ値ではないと考えるのかは、どのような観点からの「の」を研究するのか、また、それらの文書を含めると分析結果にどれぐらいの影響を与えるのかによって基準が異なる。Brown コーパスでは、「特殊な言語使用がなされている詩、話し言葉性が強い劇、会話が 50%を超える小説は対象外にする方針」が立てられていた。BCCWJ では、無作為抽出を重視する立場から、このような選別は行われていない。このため、使用者が文書の選択を判断していく必要がある。分析の目的やそれらの外れ値を含めても影響が軽微な場合は、外れ値をそのまま分析に含めることもあるだろう。ここで重要なことは、外れ値の存在を把握し、その外れ値を評価した上で分析を行う必要があるという点である。外れ値の存在に気づかないまま分析を行った結果、外れ値によって影響された結果を、調査対象の特徴と誤認することだけは避けなければならない。統合形式の「の」の頻度は、1,543,392 語であるため、用例を読むことによって、このような特殊な文書を見つけ出すことは困難である。散布図を描けば、特徴的な文書が見つけやすく、そのテキストを確認することで、その文書がなぜ外れ値となっているのかの理由を見つけやすい。

次に、頻度順位 100 位の「そこ」の散布図を観察する（図 4.11）。このグラフの場合、語数と頻度との強い相関は見られない。ピアソンの積率相関係数は.430 で、中程度の相関に留まっている。図中右端の文書は『ユリシーズ』である。「の」と異なり、「そこ」の場合、文書の語数が多くても頻度が高くなっていない。

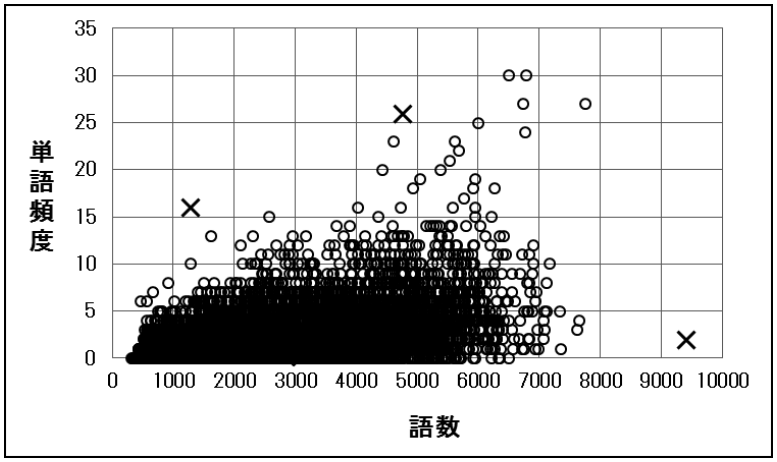


図 4.11　語数と代名詞「そこ」の頻度の散布図（×印は本文で言及する文書）

その一方で、語数が少ない割に、高い頻度で「そこ」を使用している文書がある（図4.11 左端の×印）。(11)はこの文書の一部である。

(11) どんな大きな自然に向かいあっても自分を忘れ切ることは望んでも出来ない。しかしそれの出来る場所がどうもあるらしい。そこが今言った谷の奥の広い土地である。ただそこの一切について誰も語らないのであるから、私が知っている限りのどこに似ているらしいとも何とも言えない。また仮りに、私がそこを訪れたい願いを抱いて、地上を隈なく探し歩いたあとでやっと発見したとしても、そこへ入れば、後になって語れるような感覚や記憶力を奪われてしまうのだから、結局はその努力も無駄になることは最初から分っている。

（LBj9_00024，串田孫一，『新選山のパンセ』）

この用例では、特に不自然な特徴は見当たらない。確かに「そこ」が多用されているが、名前の付けられない場所を「そこ」と呼んでおり、この文書における「そこ」は、一種の話題語になっているため、高頻度になっていると思われる。

しかし、次の(12)には問題があると思われる（図 4.11 中央の×印）。なお、(12)では「そこ」の用例は一か所に固まって出現していないため、複数の場所から引用した。

(12) ① 小学校でも、旧穢多の子弟は、本堂や拝殿の縁側に薄べりを敷いて、**そこ**で学ばせた。

② 余りに子供らしい事を習わせられるのだから、一般の者が本気で習わない。**そこ**で私はわざとその仲間へ入って他と同様に伝習を受けた。

③ この九月に東京に居る父が大病に罹って危篤だという知らせがあった。**そこ**で私は驚いて、県庁に願って東京へ赴くこととした。

（LBq9_00217，内藤鳴雪，『鳴雪自叙伝』　注：①～③の記号、および「**そこ**」の太字・下線は筆者による）

(12)で使用されている「そこで」は、①は代名詞＋格助詞の用例である。一方、②と③は、「そこで」という接続詞の用例である。BCCWJ が形態素解析に使用している辞書では、「そこで」という接続詞を認定しておらず、すべて代名詞＋格助詞に形態素解析している。このため、これらが混在して検索されてしまうのである。しかし、代名詞の「そこ」と接続詞の「そこで」は異なる単語だと認定する考え方もできる。その場合、

検索で抽出された「そこで」の用例をすべて吟味して、再分類する必要が出てくるだろう。

このような調査対象外用例のチェックは、用例検索を行った段階で、一通り用例を読むことで見つけられる場合もあるが、「そこ」の場合、用例数は 4,289 あり、用例を読んで問題点を見つけることは容易ではない。図 4.11 のような散布図を使用した分布観察を行うと、特徴的な文書が際立つため、そのような用例を集中的に調査することで、何らかの異常や言語的特徴を発見できる場合が多い。

図 4.12 は頻度順位 500 位の「社会」の散布図である。ピアソンの積率相関係数は.121 で、語数と頻度の相関は見られない。「社会」のような話題語の場合は、語数が多くなったからといって必ずしも頻度が高くなるわけではないことが分かる。最も語数が多い『ユリシーズ』（図 4.12 右端の×印）では 1 例も出現していない。

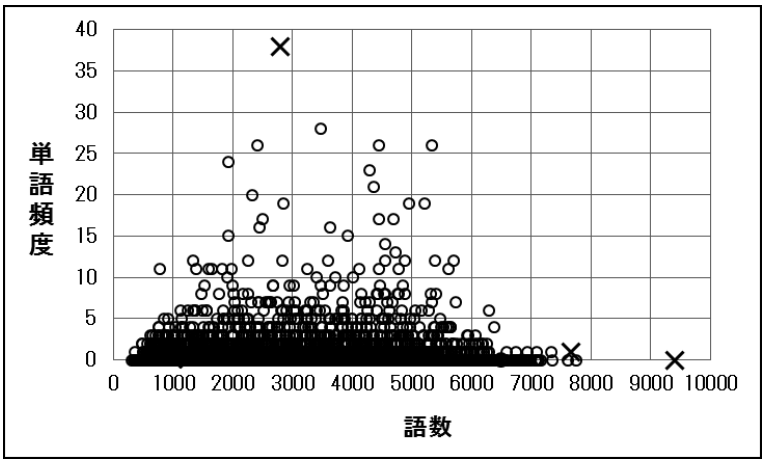


図 4.12　語数と「社会」の頻度の散布図（×印は本文で言及する文書）

(13)は最も頻度が高い文書からの引用である（図 4.12 左上の×印）。

(13) 『自由論』Ｊ・Ｓ・ミル塩尻公明・木村健康訳（岩波文庫）自由な社会のむずかしさ　自由な社会とは、矛盾に満ちた社会のことだ。　わたしがヘーゲル哲学から学んだ重要なものの見かたの一つがそれだ。矛盾を突きつけられれば、人びとはそれを解決しようと努力を重ねる。その努力があらたな矛盾をうみだす。そのようにして前へと進んでいくのが自由な社会だ。ヘーゲルはそう考えた。　（LBs1_00010，長谷川宏，『いまこそ読みたい哲学の名著』）

(13)は、哲学の名著 15 編を解説している書籍の一部である。ここではミルの『自由

論』が取り上げられている。この文書では、語数が 2,783 語の中で 38 回「社会」が出現するが、それはデータに何らかの異常があるためではなく、「社会」という単語がこの文書の主要な話題であるからだ。逆に頻度 1 で語数が最も多い文書は、林陸朗『光明皇后』（LBa2_00007）の一部（図 4.12 右下中央寄りの×印）で、話題は光明皇后が崩御する頃の宮廷や日本全体の状況である。文書の中心的話題は「社会」ではないが、「社会」という単語が複数出現してもおかしくないような話題の中で、「社会」が 1 回だけ使用されている文書である。この文書にも特に問題は見られない。

このように、「社会」という単語は、「社会」そのものが中心的な話題であれば一定量の語数の中で繰り返し出現するが、文書の語数が多いからといってそれに連動して頻度が増えるわけではないことが分かる。順位 500 位の高頻度語ではあるが、頻度 10 を超える文書はそれほど多くはなく、同じような頻度でも文書の語数は少ないものから多いものまでさまざまである。

次に、同じ話題語である頻度順位 9,890 位の「乗り物」（図 4.13）と 20,000 位の「虫歯」（図 4.14）の分布を観察する。図 4.8 と図 4.9 は第 1 節で観察した固定長を使用した度数分布図で、後で比較するために再掲している。

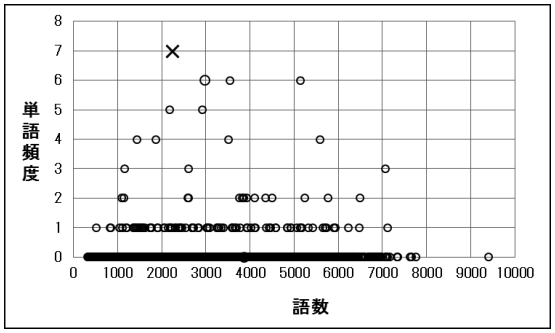


図 4.13 語数と「乗り物」の頻度の散布図

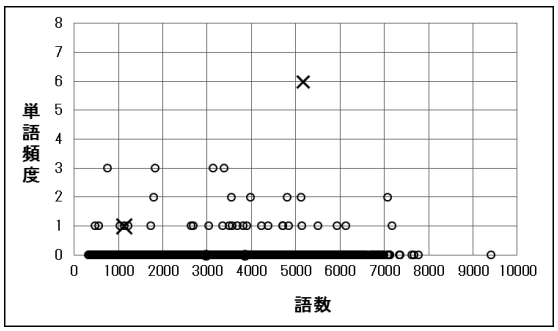


図 4.14 語数と「虫歯」の頻度の散布図

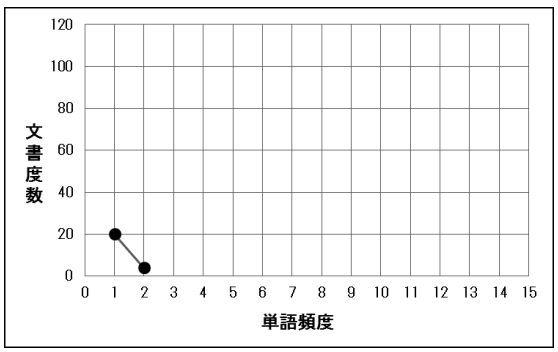


図 4.8 「乗り物」の文書度数折れ線（再掲）

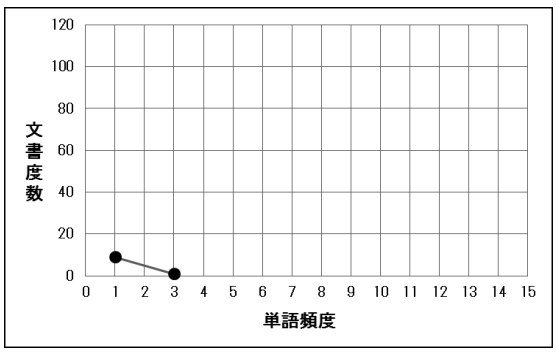


図 4.9「虫歯」の文書度数折れ線（再掲）

「乗り物」の最大頻度は 7、「虫歯」は 6 で、「社会」の 38 とは大きく異なっている。最大頻度が出現する出典も「乗り物」は、ポーラ・ゴズリング（著）山本俊子（訳）『ハロウィーンの死体』（LBo9_00115）（図 4.13 左上の×印）、「虫歯」は、藤原伊織『ひまわりの祝祭』（LBo9_00147）（図 4.14 中央上の×印）と、どちらも小説になっており、「乗り物」や「虫歯」がテーマになっている書籍から抽出された文書ではない。「社会」の場合、これが出現する文書の書名に「社会」が含まれる書籍が 48 冊あるが、「乗り物」が出現する文書の書名の中に、「乗り物」という単語が含まれている書籍は 1 冊もない。「虫歯」の場合は、志村則夫『歯医者に虫歯は治せるか』（LBl4_00001）（図 4.14 左下の×印）という書名の書籍があるが、抽出された文書の話題は留学中のストレスによって歯茎から出血した話題が中心で、虫歯はこれに関連して出現するだけで、頻度は 1 となっている。「乗り物」や「虫歯」は、「社会」に比べて図書館に所蔵されている書籍では話題になりにくいため、頻度が低いと考えられる。

固定長→統合形式による頻度の増加は「乗り物」が 28→174（6.2 倍）、「虫歯」が 12→56（4.6 倍）である。長単位の場合、固定長全体の語数は 550 万語強、統合形式全体の語数は約 2,500 万語で、統合形式の語数は固定長のおよそ 4.5 倍になっている。これに連動して、高頻度語では固定長頻度に比べ統合形式頻度が 4.5 倍前後になっているものが多く、「の」が 4.5 倍、「そこ」が 4.7 倍、「社会」が 4.5 倍という倍率になっている。しかし、「虫歯」が 4.6 倍と同程度であるのに対し、「乗り物」は 6.2 倍と大きく倍率がずれている。

「乗り物」と「虫歯」の散布図と固定長の文書度数分布折れ線を比較すると、図 4.13 と図 4.14 の散布図では「乗り物」の方が頻度 4 以上の文書が多いのに、図 4.8 と図 4.9 の文書度数折れ線では、「乗り物」に頻度 2 の文書しかない。逆に散布図では頻度 3 以上の文書が少ない「虫歯」の文書度数折れ線に頻度 3 の文書が見られる。本来であれば、図 4.8 の文書度数折れ線にもっと頻度 3 以上の文書が出現するべきで、そのサンプリング誤差のために頻度の倍率が 6.2 倍になっていると思われる。このことは、「虫歯」の固定長頻度が母集団の過不足ない縮図にはなっていない可能性を示唆している（これについては第 3 節で検討する）。

最後に頻度順位 3,010 位の「興味深い」の散布図を観察する。「興味深い」では固定長→統合形式による頻度の増加は、114→632 と 5.5 倍になっている。「興味深い」でも頻度の増加はデータ量の増加の 4.5 倍より大きいため、固定長の分布は母集団の過不足ない縮図にはなっていない可能性がある。図 4.7 の固定長の度数折れ線では、頻度 1 の

文書しか観察されず、「興味深い」という形容詞は 500 語程度の範囲内では何回も繰り返されない単語だと考えられた。しかし、図 4.15 の散布図を見ると、複数回使用する文書も多く見られる。これらは 1,000 語以上の文書で現れるため、単純に割り算をすれば 500 語に 1 回以上となるが、実際の用例ではもっと近い範囲で繰り返される例が見られる。

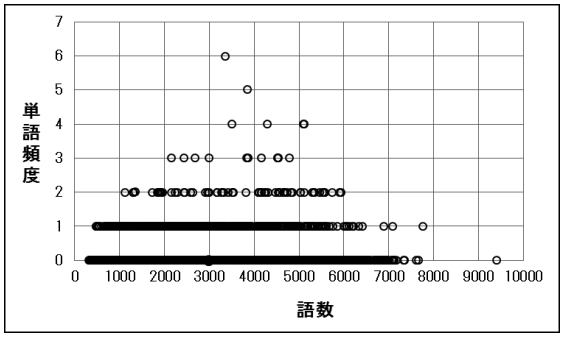


図 4.15 語数と「興味深い」の散布図

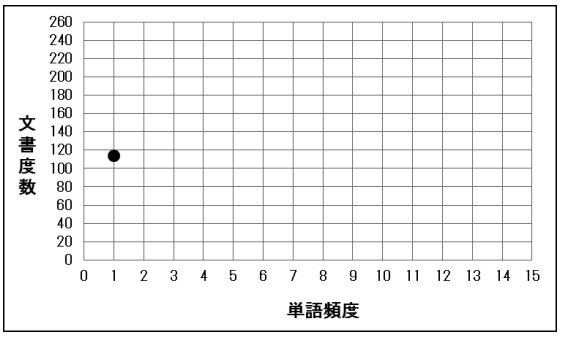


図 4.7「興味深い」の文書度数折れ線（再掲）

短い範囲で繰り返しが起きるのは、(14)のように前の表現を受け、その関連でもう一度使用される例が多い。しかし、(15)のように、異なる対象に対して繰り返し使用する例も存在する。

(14) これは、大変興味深い発見である。なぜ興味深いのか、私のエマルジョン体験から明らかにしてみたい。　（LBq7_00022，歌田眞介，『油絵を解剖する』）

(15) それは『とはずがたり』後半の志向に似通っているのも興味ふかい。それにしても、持明院への行幸に三位中将となった実俊が剣璽の役として勤める、その進退作法を見つめる名子のまなざしは、まさしく嘗ての内侍のそれに他ならない。二十一　西園寺公宗の悲劇については、『太平記』巻第十三「北山殿御隠謀事」に詳しく物語られる。それが丁度『竹向きが記』に伏せられた間の消息を補う形であることが注意され、また琵琶秘曲伝授説話の面を含むことも興味深い。　（LBg9_00073，阿部泰郎，『日本文学史を読む』）

図 4.15 の散布図や、(14)、(15)の用例を観察すると、図 4.7 の文書度数折れ線でも頻度 2 以上の文書が何例か出現してもよかったのではないかと思われる。

以上、BCCWJ 図書館 SC・統合形式を使用して、語数と単語頻度の散布図を観察した。

語数と単語頻度の散布図を使用すると、第 1 節の固定長を使用した文書度数分布の観察と同様に、外れ値となっている文書が見つけやすく、データクリーニングや出現傾向の観察が効率的に行えることが確認された。また、統合形式では頻度が固定長のおよそ 4.5 倍に増えるため、特に低頻度語において固定長より正確な分析ができる可能性が示唆された。

第 3 節　文書内の単語分布の観察

前節の観察では「乗り物」や「興味深い」において、統合形式に比べて固定長の頻度がやや少なく、固定長の分布は母集団における出現状況を正しく反映していない可能性が考えられた。本節ではなぜそのような現象が起きるのかを、文書内の単語分布を観察することで考察する。これによって、統合形式と固定長の関係も明確になると思われる。

固定長の頻度が母集団の分布を必ずしも正確に反映しない原因は、サンプルの抽出方法にある。Brown コーパスでは、出版物のリストからはじめに 500 の出版物が無作為抽出され、次にその出版物から一部の文書が無作為抽出されていた。BCCWJ の書籍コーパスでは、基本的に母集団の文字リストから直接書籍の 1 文字が無作為抽出されているが、作業を簡略化するために Brown コーパスと同様の抽出方法で集積した文書もあり、実質的には Brown コーパスと同じ方法で無作為抽出されたデータと見なすことが可能である。はじめの無作為抽出はさまざまな文書タイプを母集団と同じ比率で抽出するための行為で、無作為抽出の意味としてはこちらの方が大きい。二回目に行われる無作為抽出は、書籍全体をデータ化できないために、一部の文書で書籍全体を代表させるための抽出である。

先に述べたように、石田・佐藤（2010）によれば、1,000 語のデータよりは 5,000 語のデータの方が精度が高く、同じ 1,000 語であれば 1 か所から連続して取得するより、200 語ずつを 5 か所に分散して取得した方が精度が高くなる。この結果は、固定長より統合形式の方が精度が高いデータであることを示唆している。その理由は単純で、1 冊の書籍の言語的特徴を代表するには、短い文書より長い文書の方が偶然による誤差が少なくなるからである。

そこで、書籍から一部の文書を抽出することによってどれぐらいの誤差が生じているかを検証するため、本節では、統合形式と固定長の文書内の単語分布を比較する観察を行う。観察の対象は前節までの観察でも使用した、BCCWJ 図書館 SC・固定長・長単位の語彙頻度表で順位 1 位の「の」、100 位の「そこ」、500 位の「社会」、3,010 位の「興

味深い」、9,890 位の「乗り物」とする。

文書内の単語分布を観察するには、調査対象が文書内のどの位置に出現するかを特定する必要がある。これには国立国語研究所が Web 上で公開している検索アプリケーション「中納言」の検索データが利用できる。「中納言」を使用して BCCWJ から調査対象を検索したデータには、用例以外にもサンプル ID、品詞、書名などさまざまな情報が付与されている。そのような情報の一つに、その調査対象が統合形式のサンプルの先頭から何番目の短単位に当たるかを表す「連番」という情報がある。この「連番」を使用すると、ある文書の中で、調査対象がどのように分布しているかを描くことができる。「連番」は記号等を含めた短単位の順番を表しているため、以下の調査は短単位データをもとに行う。短単位の順番が 750 番の場合、その区間には記号を含めた単語が 750 語含まれていることを意味するため、*x* 軸のラベルは語数とした。

図 4.16 は、文書内で、頻度順位 1 位の「の」がどのように分布しているのかを描いた図である。一つの書名に対して上下二つの散布図が対応しており、上段が固定長、下段が統合形式での単語の分布を表している。下段に記した統合形式の右側に×印で描いているのが統合形式の末尾である。固定長の先頭と末尾の位置は描いていないが、記号等を含めた短単位の平均的な長さは 750 語であるため、750 語刻みの *x* 軸の補助線が一定の目安になる。

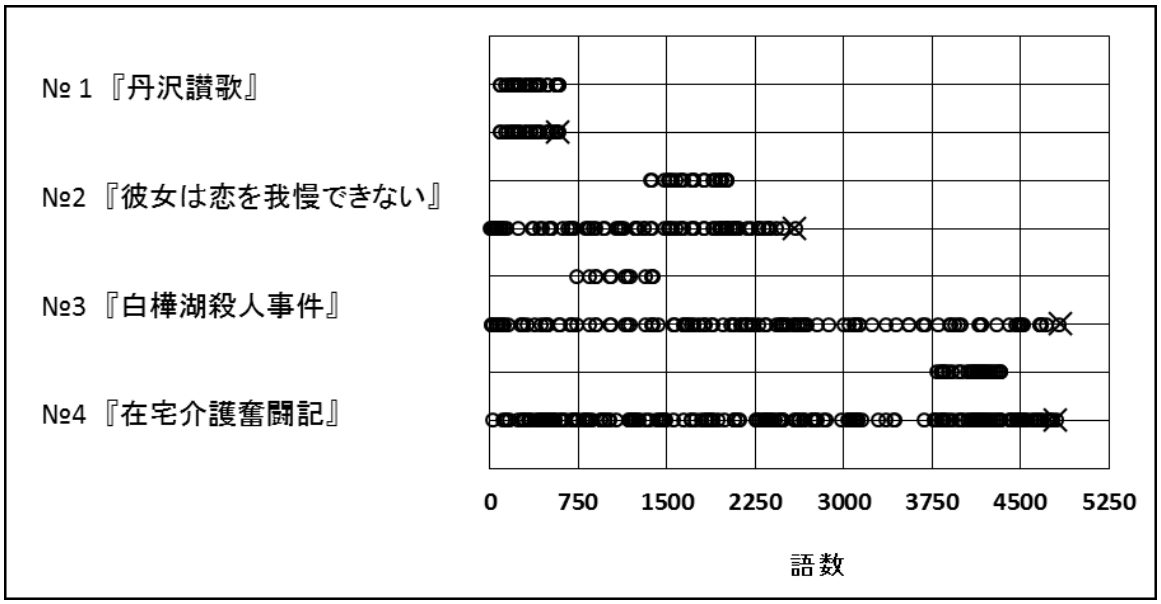


図 4.16 固定長の単語分布（上）と統合形式の単語分布（下）:「の」

文書は短い文書から長い文書までランダムに選んだが、「の」の場合、ほとんど文書

全体にまんべんなく出現するため、これらのどの部分から固定長を抽出しても、固定長は統合形式のよい縮図となる。これは統合形式に限らず、1 冊の書籍全体でも同じような分布となっていると考えられるため、「の」の場合、固定長は高い精度で書籍を代表するサンプルになっていると思われる。小規模なコーパスしか製作できなかったコーパス言語学の初期から、高頻度の言語現象であれば高い精度の分析ができるといわれてきたのは、高頻度語の場合、このように少数のデータでも誤差が生じにくいサンプリングができるからだと考えられる。

図 4.17 は、同様の方法で頻度順位 100 位の「そこ」の文書内分布を描いた図である。「そこ」になると、「の」とは異なり、文書の全体にまんべんなく出現するわけではない。この場合、どの位置から固定長が抽出されるかで誤差が生じる。たとえば№3『大航海時代』の固定長頻度は 0 であるため、統合形式で「そこ」が出現していない部分から固定長が抽出されていることが分かる。このように頻度順位 100 位の高頻度語であっても、文書全体にまんべんなく出現しない単語では、ある程度の抽出誤差は避けられなくなってくる。

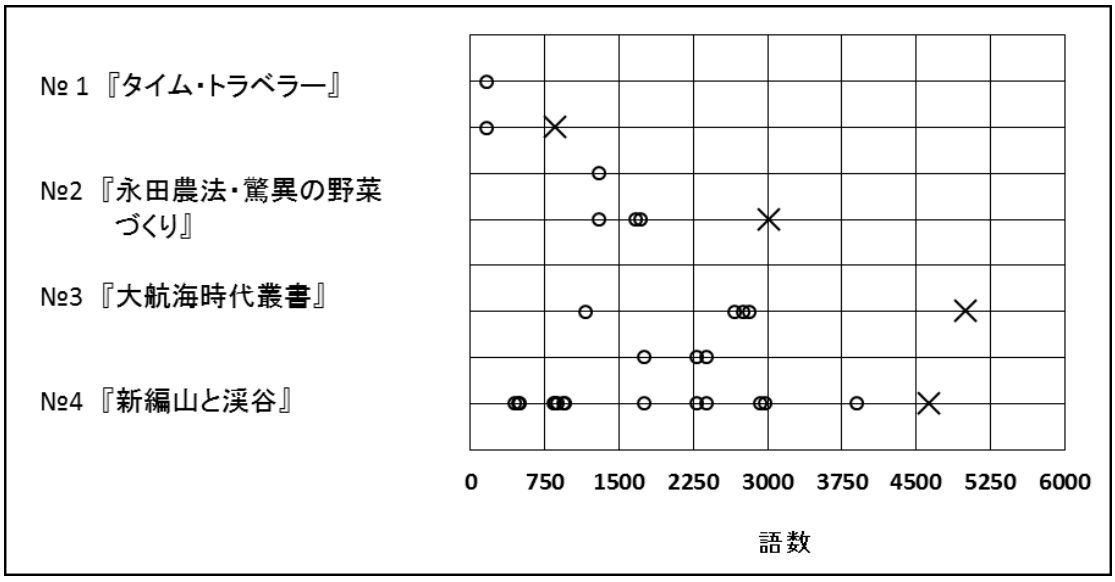


図 4.17 固定長の単語分布（上）と統合形式の単語分布（下）:「そこ」

これが話題語になれば、固定長が母集団の正確な縮図になっている確率がさらに低くなるのは当然である。図 4.18 は、頻度順位 500 位の「社会」の分布を描いた図である。前節で観察したように話題語では語数と単語頻度との連動性がなくなるため、これ以後は固定長の頻度が高い順に、出現状況を調査した。図 4.18 では、固定長の頻度が高い

順に文書を選んでいることもあるが、すべての文書で「社会」が頻出する箇所から固定長が抽出されていて、固定長はかならずしも統合形式のよい縮図になっているわけではない。

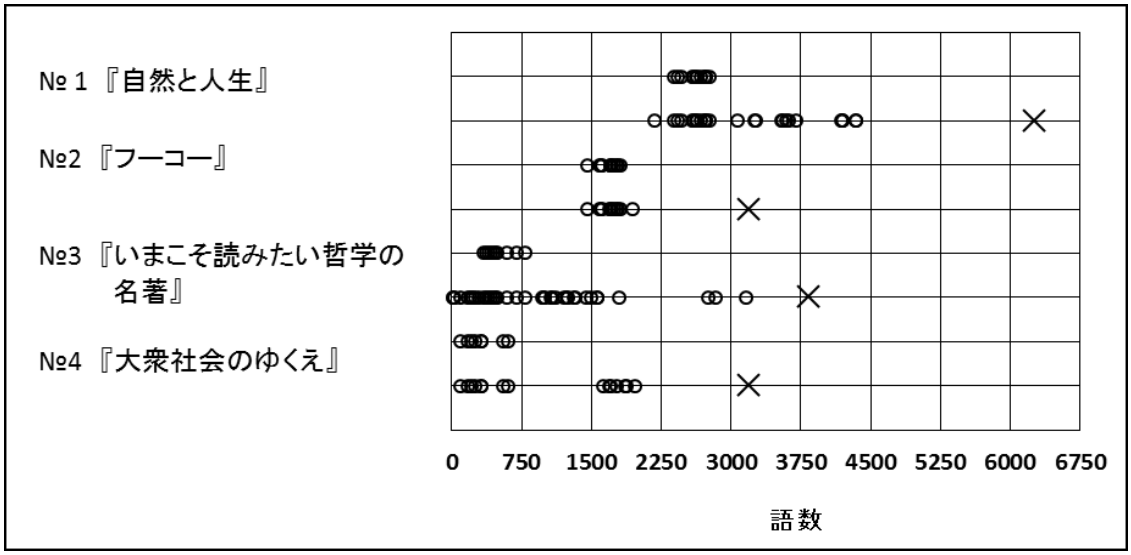


図 4.18 固定長の単語分布（上）と統合形式の単語分布（下）:「社会」

もちろん「社会」が出現する文書の全てを観察すれば、統合形式では出現しているのに、固定長では出現していない文書もあって、それらをすべて合計して計算すれば固定長の平均は母平均と近似している可能性がある。標本の大きさが大きくなるにつれて、標本平均は母平均に近づくことが知られており、これを大数の法則という（Upton & Cook，2011:230-1；尾畑，2014:147-51；小島，2006:130 など）。「社会」が出現した固定長の文書数は、742 であるから、大数の法則が働くことも考えられる（どれぐらいの出現数があれば母集団に対して誤差 5%以内の分布となっていると考えられるかについては、第 5 節で分析を行う）。

図 4.19 は頻度順位 3,010 位の「興味深い」の分布図である。この図では統合形式の頻度が 4 以上の六つの文書の分布を描いた。文書数を増やしたことによって固定長と統合形式の違いが見にくくなってしまったため、固定長の出現位置は◆で示している。図 4.19 では、6 文書のうち固定長に出現しているのは、№3 と№6 の 2 文書だけで、それぞれの固定長頻度は 1 になっている。この 6 文書で見ると、統合形式の頻度に比べ、固定長の頻度は明らかに少ない。「社会」のように固定長で 742 文書に出現する単語であれば、互いの誤差を打ち消し合ってデータ全体では正確な頻度となることが期待できるかも知れない。しかし、「興味深い」のように固定長での出現文書数が 114 の単語では、

合計や平均値を求めても、抽出の偏りがそのまま残ってしまう可能性がある。

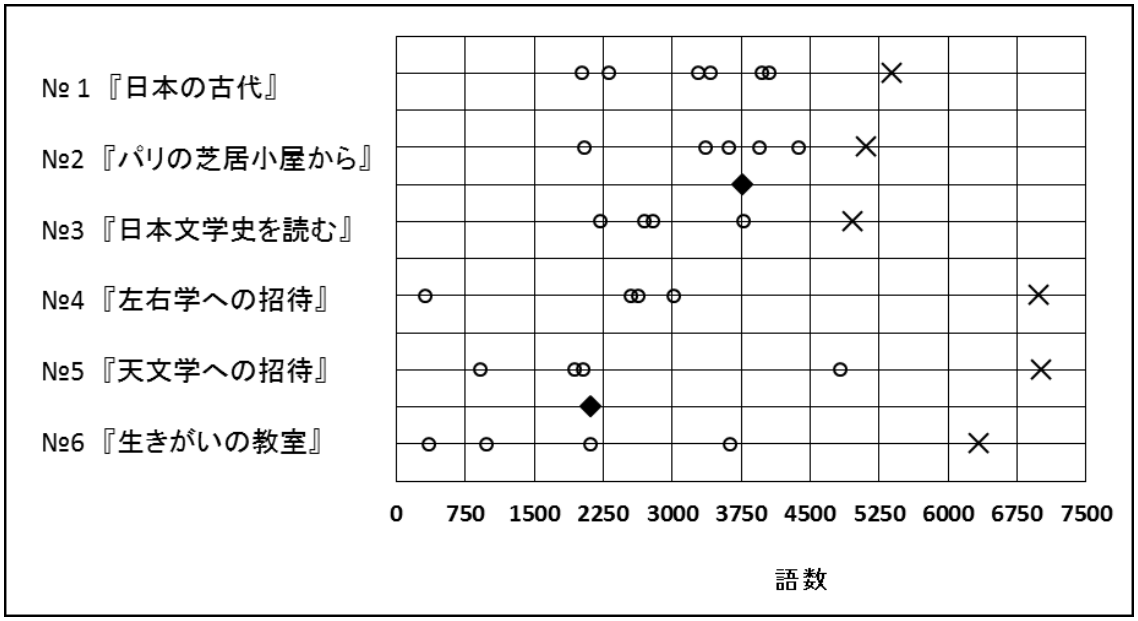


図 4.19 固定長の単語分布（上）と統合形式の単語分布（下）:「興味深い」

なお、図 4.19 のような文書内の単語分布を描けば、第 2 節で検討した「興味深い」という形容詞がどれぐらいの間隔で使用されるのかという疑問を直接観察することができる。この図の中でマーカーが連続して出現する№1、3、4、5 の用例は、(13)で観察したような、一度使った「興味深い」という表現を受け、その関連でもう一度「興味深い」が使用される例である。しかし、これ以外の文書では用例(15)のように「A は興味深い」「B も興味深い」と、同じ文型で使用されている例も見られる。

固定長を使用した度数折れ線（図 4.7）では、頻度 1 の文書しか観察されないため、「興味深い」という形容詞は、短い範囲での繰り返しが避けられている単語だと考えたくなるが、実は、図 4.7 の結果はサンプリングの偏りから生じた結果であり、語数を増やした文書で観察すると 500 語の範囲であれば繰り返し使用されるケースも少なくないことが分かる。

最後に図 4.20 で「乗り物」の統合形式頻度が 5 以上の六つの文書を描いて観察する。この 6 文書では、固定長に出現する文書が 3 文書で、№1、5、6 に出現している。№3 や 4 では、短い範囲に集中して出現しているため、出現箇所が固定長と重なりにくかったと思われる。この図でも統合形式の○の数に比べ、◆の数が少ないのは明らかである。「乗り物」は固定長の文書数が 24 であり、大数の法則は働きづらいと考えられる。

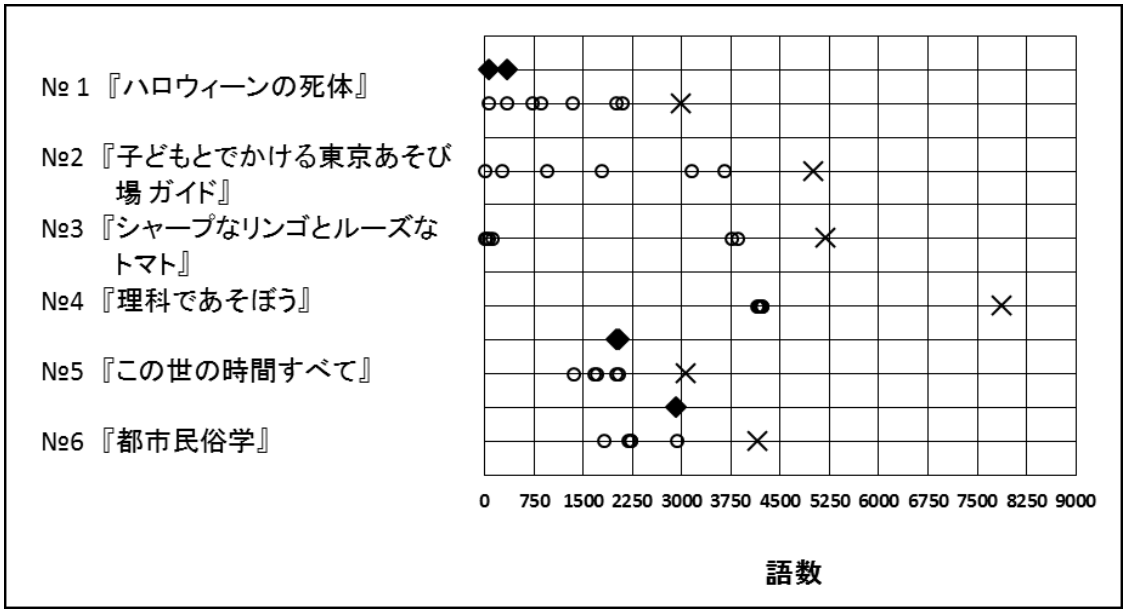


図 4.20 固定長の単語分布（上）と統合形式の単語分布（下）:「乗り物」

以上、文書内の単語分布を観察し、固定長がどれほど統合形式のよい縮図となっているかを観察した。「の」のように、文書の全体にほぼまんべんなく出現する単語の場合は、母集団の分布に近いサンプルになっていると考えられる。しかし、頻度順位 100 位の「そこ」や 500 位の「社会」という高頻度語であっても、一つ一つの文書では誤差を持っていることが考えられた。この誤差は、出現する文書数が多ければ、大数の法則が働いてその平均値は母平均に近づくと考えられるが、文書数が少なくなると、偏った分布を持つ可能性が増加する。統合形式は固定長の 4.5 倍のデータ量があるため、1 文書当たりの語数が多く、調査対象の言語現象が出現する文書数も多い。このため、特に低頻度語では、固定長より統合形式の方がより母集団の分布に近い分布を示していると考えられる。

第 4 節　文書の語数が異なるコーパスでの文書度数分布の観察法

本節では、文書ごとに異なる語数を平準化して文書度数分布図を描く方法を検討する。BCCWJ の場合、固定長より統合形式の方が母集団のよりよい縮図となっている可能性が高い。しかし、統合形式がこれまでほとんど統計分析に使用されてこなかった理由は、一つ一つの文書の語数が異なっているためである。文書の語数が異なると、同じ条件で頻度比較を行うことができず、分析に使いづらい。広義コーパスでは、統合形式のように個々の文書の語数が異なっていることが多い。このため、本節の方法は第 1 節の方法

より汎用性が高い文書度数分布の観察法となる。第 4.1 項では、個々の文書の語数を平準化する個別調整頻度の算出方法について説明する。第 4.2 項では個別調整頻度を使用して度数折れ線を観察する方法を検討する。

第 4.1 項　個別調整頻度の算出方法

先行研究において、語数が異なるコーパスを比較する方法として使用されてきたのが調整頻度（換算頻度, 正規化頻度）である（バイバー・コンラッド・レッペン, 2003:38-41；石川・前田・山崎，2010:27-8；石川，2012:114-5；マケナリー・ハーディー，2014:74-6 など）。調整頻度はあまりにも一般化しているためか、丁寧な説明が書かれている文献はそれほど多くない。Gries（2010:271）では「While this is so basic as to hardly merit discussion, it is a point of entry for some other concepts.（筆者訳：これはあまりにも基本的であるため、議論のメリットがほとんどないが、他の概念の入り口となっている）」と記されている。次の引用は、石川（2012:114-5）で記述されている調整頻度の説明である。

> 調整頻度（adjusted frequency）とは、異なるコーパスから得られた頻度の相互比較が可能になるよう、粗頻度を一定の基準によって調整した値のことです。〔･･･〕。
>
> すでに述べたように、粗頻度のままでは複数のコーパスから得られた頻度を比較することはできません。たとえば、ある語の頻度が 500 万語コーパスで 10 回、2,000 万語コーパスで 30 回だったとして、単純に 10 と 30 を比較することに意味はありません。
>
> この場合、共通の物差しを用意して、個々の頻度をそれに合わせて調整する必要があります。頻度の調整にはさまざまな方法があり、最も単純な方法は、観測された度数を総度数で割り、相対頻度（relative frequency）とすることです。〔･･･〕。
>
> ただし、これでは、あまりに値が小さくなりすぎ、研究には不便です。そこで、コーパス言語学では、単純な相対頻度に代えて、任意の調整基準を定め、基準に対する比率値として粗頻度を調整します。このようにして作られた値を調整頻度と呼びます。
>
> 〔･･･〕調整頻度の基準値にはさまざまな値を使用できますが、値をある程

度大きくして読み取りやすくするため、コーパス研究では、通例、「100 万語当たり調整頻度」が使用されます。「100 万語当たり」というのは英語では per million words と書きますので、頭文字を取って PMW と表記します。〔･･･〕。

なお、PMW は最も一般的な基準値ですが、どんな場合でも PMW が適当なわけではありません。たとえば、5 万語のコーパスと 20 万語のコーパスの頻度を PMW でそろえるのは不適とする立場もあります。実際には 100 万語まで調べていないのに、あたかも 100 万語を調べたように見えてしまうためです。一般に、元のサイズを超えて基準値を設定することはデータが本来持つ情報量を過大解釈することにつながるため、好ましくありません。(石川, 2012:114-5)

この引用からも分かるとおり、調整頻度とは相対頻度に一定数をかけた頻度のことで、これよりさらに基本的な統計量は相対頻度である。相対頻度はある単語がコーパス全体の何%を占めているかという割合を表しているため、「使用率」と呼ばれることもある。この相対頻度がいつから言語分析に使用されることになったのかは不明だが、アメリカ心理学協会の学会誌第 1 号には、すでにこの相対頻度を使用した論文が掲載されている。

表 4.1　生後 19 か月男児の発話の品詞割合

Dewey（1894:64）より引用

A at 19 *mos. old.*

Parts of Speech.		Per cent.
Nouns . . .	68	60
Verbs . . .	24	21
Adjectives .	13	11
Adverbs . .	4	3
Interjections	6	5
Total . .	115	100

Pronouns, prepositions, conjunctions, none.

Dewey（1894）は児童の言語の発達を扱った研究で、表 4.1 では、A という生後 19 か月の男児が発話した 115 語の中で各品詞が何語あったかの粗頻度とともに、相対頻度（Percent）が記述されている。この論文では 20 か月の女児のデータと合わせた 2 人分の品詞の相対頻度を、先行研究のデータと比較して論じている。これを見ると、データ量が異なる言語データ同士を相対頻度に直して分析することは、120 年以上前から行われていたことが分かる。

調整頻度は相対頻度に一定数をかけただけの値なので、「議論のメリットがほとんどない」といわれるほど基礎的な統計量だといえる。ただし、この調整頻度でどれぐらい正確に調整できるかは、単語の性質によってかなり異なる。第 2 節で観察した頻度順位 1 位の「の」のように、語数に連動して頻度が増える単語であれば、かなり正確に調整できる。しかし、100 位の「そこ」の場合、語数と単語頻度の積率相関係数は.430、500 位の「社会」では.121 と、頻度が下がるにつれ、語数と頻度の相関は急激に失われる。低頻度の単語の場合、調整頻度による平準化は不正確な値になることが考えられる。

かといって、語数の異なるデータを比較しようとした場合、この調整頻度以外にこれといってよい方法があるわけではない。このあたりの事情について、木村（1982:226）では、「むろん標本使用率と母使用率とが完全に一致するものではないことはだれしも承知しているが、といって他にどうしようがあるかと言いたいところである」と述べられている。これは多くの研究者にとって共通の思いであろう。

石川（2012:114-5）の説明は、コーパス全体で集約した頻度の調整について述べられているが、本節の目的は文書ごとの頻度分布を観察することにあるため、文書ごとに調整頻度を算出する必要がある。つまり、以下の式が個々の文書の調整頻度の算出式で、本研究ではこれを「個別調整頻度」と呼び分ける。単に調整頻度といった場合は、従来の先行研究で使用されているとおり、コーパスや出版媒体、学習者の習得レベルなど、何らかのカテゴリーで合計して平準化した場合の調整頻度を意味する。

（4.1 式）　　個別調整頻度　＝　文書の観測値　÷　文書の語数　×　基準値

第 4.2 項　個別調整頻度を使用した文書度数分布の観察法

調整頻度を算出する際、問題になるのが何語当たりの調整頻度にすればよいかという基準値の値である。石川（2012:115）では「一般に、元のサイズを超えて基準値を設定することはデータが本来持つ情報量を過大解釈することにつながるため、好ましくありません」とあった。しかし、たとえば 5 万語のコーパスと 50 万語のコーパスを比較する際、5 万語のコーパスに合わせるということは、50 万語のコーパスを 5 万語の精度に引き下げて観察することにもなりかねない。調整頻度とは相対頻度に一定数をかけたものであるから、たとえば相対頻度が 0.0000246 の場合、これを 5 万倍にすると 1.23 語、50 万倍にすると 12.3 語となる。1.23 語では値が小さすぎて小数点以下の値を評価しにくい。1.23 語と 1.45 語の違いはほとんどないように感じるが、12.3 語と 14.5 語では違

いを感じやすい。つまり、基準値をどれぐらいの値にするかは、有効桁数をどれぐらいに設定するかという問題と類似しており、調整した頻度のどのあたりから誤差と見なすかにかかっている。

BCCWJ 図書館 SC の固定長と統合形式における、文書ごとの語数分布は図 4.21 のようになっている。図 4.21 で固定長の語数が太い線になって見えるのは、*x* 軸を統合形式の順に並べているため、固定長ではばらつきが生じてマーカーが上下にばらついて並ぶためである。統合形式語数の最小値は 313 語、最大値 9,406 語、平均 2,378 語、標準偏差 1,466 語、変動係数 0.62 である。

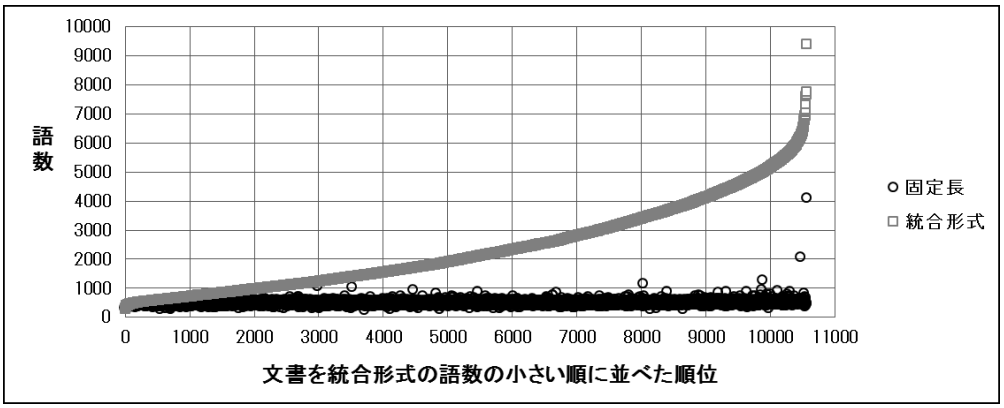


図 4.21　図書館 SC・長単位・固定長と統合形式の文書別語数分布

1 冊の書籍が持っている言語的な特徴を、一部の文書で代表させる場合、語数が多いほどその特徴をよりよく反映できる。この文書を固定長平均に近い 500 語当たりで調整した場合、頻度合計自体は固定長とそれほど変わらないが、一つ一つの文書がもとの文書のよりよい縮図となっている割合は高くなる。文書によって精度の高さにはばらつきがあるが、そのばらつきによってデメリットが生じるとは考えにくい。最も精度が低い文書で、固定長と同じ精度になっている。

しかし、その精度の高さを、具体的に分析に生かせるかどうかはまた別問題である。500 語当たりで調整した場合、精度の高さは小数点以下の値となって表れるが、これを生かした状態で文書度数折れ線を描くことは難しい。個別調整頻度を四捨五入して整数に直すとすると、せっかく精度の高い情報を得ておきながら、その多くを捨てることになる。小数点以下の情報を捨てない方法として、小数点以下を繰り上げて整数に直すことが考えられる。0 から 1 までの区間に落ちた値をすべて 1 の階級に数えるのである。この場合、1 の頻度が非常に高くなる。固定長と統合形式のデータを比較するのではな

く、統合形式単独の分布観察が目的であれば、語数を固定長平均に合わせる必要はないため、統合形式平均の 2,378 語で調整する方法も考えられる。この場合、統合形式を使用することによって増加した情報量が頻度 1 に集中せず、細かな分布が観測できる。ただし、今度は平均語数より少ない文書の頻度を過大に調整してしまうという問題が生じる。基準値を平均値の 2 倍に近い 5,000 語にした場合、過大調整が多くの文書で起こるため、石川（2012:115）が指摘するような「過大解釈」が起きやすい。基準値をどのような値にするかは、重要な検討問題である。

これらのどの基準値を使用すれば有益な分析ができるのかを検討するため、以下で頻度順位 100 位の「そこ」、500 位の「社会」、3,010 位の「興味深い」で文書度数折れ線を描き、比較を行う。図 4.22 は、固定長のデータを使用して描いた「そこ」の文書度数折れ線、図 4.23 は統合形式を使用し、調整頻度の基準値を 500 語にした上で、小数点以下を四捨五入して描いた文書度数折れ線である。この二つは似た分布を示し、固定長と統合形式との比較には向くが、統合形式が持っている情報量は生かされていない。

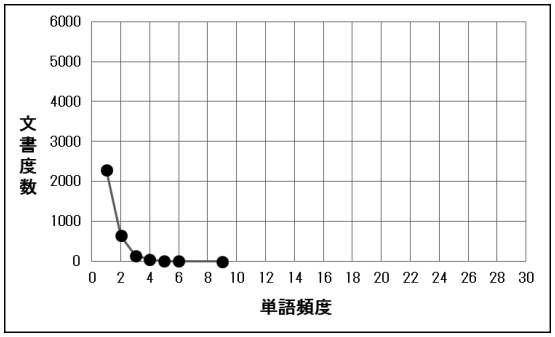


図 4.22 「そこ」固定長

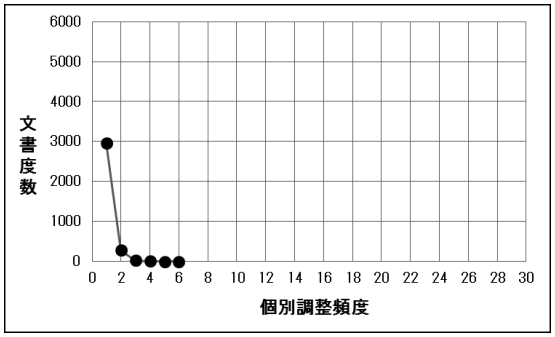


図 4.23 「そこ」統合形式：500 語・四捨五入

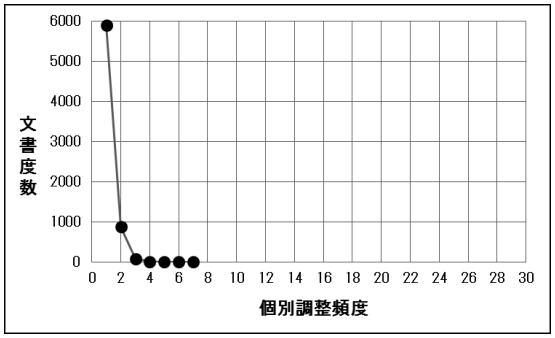


図 4.24 「そこ」統合形式：500 語・繰り上げ

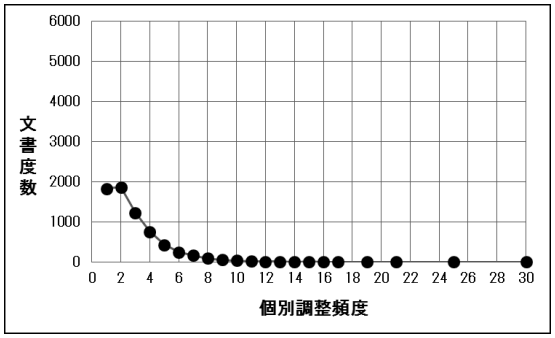
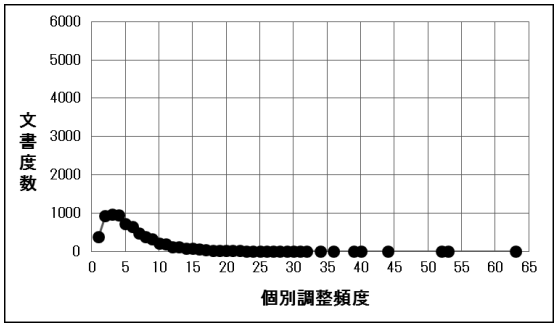


図 4.25 「そこ」統合形式：平均・四捨五入

図 4.24 は基準値を同じ 500 語にして、小数点以下を繰り上げた図である。この場合、

統合形式によって増加した情報のほとんどが頻度 1 に集中し、情報の切り捨てはないものの、細かな情報量は増えていない。図 4.25 は、基準値を統合形式平均の 2,378 語にし、小数点以下を四捨五入して描いた図である。図 4.24 のように増えた情報が頻度 1 に集中することなく、他の頻度にも分かれて表現されている。ただし、今度は頻度 20 を超えるような文書が複数出現しており、粗頻度が過大調整されていることが考えられる。



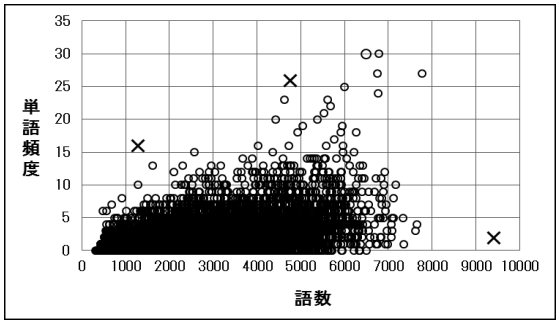


図 4.26 「そこ」統合形式：5,000 語・四捨五入　図 4.11「そこ」統合形式の散布図（再掲）

図 4.26 は、5,000 語で調整したグラフである。これを見ると非常に多くの文書が過大調整されており、図 4.11 の散布図と比べてもあまりにも実態と異なっている。図 4.11 の散布図で語数 5,000 語付近の単語頻度を見ると、最大頻度は 25 程度である。しかし、図 4.26 では単語頻度が 25 を超える文書が多数出現している。ここまで来ると、過大調整によるデメリットの方が多く、この度数折れ線を観察しても有効な分析にはなりにくい。

以上、統合形式の語数を平準化する基準値として、500 語・小数点以下四捨五入、500 語・小数点以下繰り上げ、統合形式の平均（2,378 語）・四捨五入、5,000 語・四捨五入という 4 つの値を使用した場合、文書度数折れ線がどのように変化するかを比較した。この比較からすると、固定長と統合形式のデータを比較する場合は固定長平均の 500 語・四捨五入が向いているが、統合形式単独の分布観察が目的であれば、統合形式の平均（2,378 語）を基準値とする方法が妥当だと思われる。しかし、統合形式の平均で調整した場合、一部の文書で過大調整が起きている。そこで、この過大調整によってどの程度の影響が生じるのかを、以下で詳しく観察する。

表 4.2 は、統合形式平均で調整した「そこ」の個別調整頻度を、頻度の高い順に 20 文書並べ、これらの書名、粗頻度、文書の語数を併記した一覧表である。図 4.25 で見ると、単語頻度が 13 以上の文書を、細かく観察していることになる。これらの文書の語数は、1,000 語以下が多く、平均語数より少ない文書の粗頻度を過大に調整している

ことが分かる。ただし、個別調整頻度が 15 前後の文書であれば、600～800 語の文書に 4、5 回の出現であるため、図 4.11 の散布図と比べてもそれほど過大調整になっているわけではない。

表 4.2 「そこ」の統合形式：平均を基準値とした個別調整頻度上位 20 文書の書名と語数

ID	執筆者	書名	粗頻度	語数	個別調整頻度
LBgn_00018	河野　一郎	英語の詩	6	473	30
LBj9_00024	串田　孫一	新選山のパンセ	16	1274	30
LBa3_00048	実著者不明/関　楠生(訳)	世界の民話	7	672	25
LBg4_00010	市川　平三郎	百歳まで生き、ガンで死のう。	6	567	25
LBp3_00079	神崎　宣武	三三九度	8	915	21
LBl9_00058	夢枕　獏	涅槃の王	13	1613	19
LBq1_00010	秋月　龍珉	無門関を読む	10	1278	19
LBpn_00007	実著者不明	少女が運んだ中国民話	4	559	17
LBf1_00015	実著者不明/関根　正雄(訳)	旧約聖書創世記	5	741	16
LBb9_00124	山岡　荘八	徳川家光	4	632	15
LBg3_00022	高野　孟	世紀末地球市民革命	5	792	15
LBo2_00092	岡崎　久彦/渡部　昇一	国のつくり方	4	629	15
LBs7_00045	金子　正輝	麻雀必勝のテクニック	5	789	15
LBs9_00170	山口　瞳	人生論手帖	5	817	15
LBt3_00153	島田　裕巳	「厄年」はある！	5	816	15
LBd5_00022	植田　暁/金原　孝興	水辺都市	15	2575	14
LBq7_00062	土屋　嘉男	魚はゆらゆらと空を見る	12	2101	14
LBqn_00012	ヴァージニア・ハミルトン/金関　寿夫(訳)	人間だって空を飛べる	6	1012	14
LBd8_00010	西岡　光秋	実践・心にひびく文章の書き方	3	538	13
LBe9_00183	南條　範夫	剣士流転	4	732	13

しかし、表 4.2 の先頭の『英語の詩』は、文書の語数が 473 語と短いのに、粗頻度で 6 回出現するため、個別調整頻度は 30 となり、過大調整の疑いが濃い。(16)がその文書の一部である。

(16) ８　ＯＬＤ　ＭＯＴＨＥＲ　ＨＵＢＢＡＲＤ―Ｍｏｔｈｅｒ　Ｇｏｏｓｅ

《ハバードおばさん》ハバードおばさん戸棚とこへ行ってみた犬に骨でもあげようとところが戸棚に行ってみるとなかはからっぽすっからかんあわれな犬にはなんにもないそこでおばさんパン屋へ行った犬にパンを買うためにところが家へ帰ってみるとあわれな犬は死んでいたそこでおばさん建具屋へ行った犬に棺桶買うためにところが家へ帰ってみるとあわれな犬は笑ってた

（LBgn_00018，河野一郎，『英語の詩』）

この用例はマザーグースの訳である。「そこで」が何度も出現するのは、詩の調子やリズムを整えるためであり、元の書籍の全体でこの頻度が続くことは考えにくい。また、表の 2 番目の『新選山のパンセ』は、先に用例(11)として引用した文書で、自分の存在

を忘れきってしまえるような大自然の場所を「そこ」と呼び、話題語のように使用していたため頻度が高くなった文書であった。この文書も抽出範囲を広げてサンプリングした場合、同じような割合で「そこ」が繰り返されているとは考えにくい。

つまり、粗頻度の最大値付近の文書を調整した頻度が、現実に存在する可能性が高い個別調整頻度の最大値であり、それを超えるような頻度は、語数の少ない文書でたまたま多くの回数が使われていた用例を過大に調整しているケースが多いと考えられる。

ただし、このような過大調整を避けるために、図 4.24 のような 500 語で調整する方法を選択した方がよいとは限らない。過大調整は粗頻度を相対頻度に換算した段階で発生しており、小さな基準値で調整すればその弊害が見えにくくなるだけである。

表 4.3 代名詞「そこ」の 500 語当たりの個別調整頻度上位 10 文書の書名と語数

ID	執筆者	書名	粗頻度	語数	個別調整頻度
LBgn_00018	河野　一郎	英語の詩	6	473	7
LBj9_00024	串田　孫一	新選山のパンセ	16	1274	7
LBa3_00048	実著者不明/関　楠生(訳)	世界の民話	7	672	6
LBg4_00010	市川　平三郎	百歳まで生き、ガンで死のう。	6	567	6
LBp3_00079	神崎　宣武	三三九度	8	915	5
LBl9_00058	夢枕　獏	涅槃の王	13	1613	5
LBf1_00015	実著者不明/関根　正雄(訳)	旧約聖書創世記	5	741	4
LBb9_00124	山岡　荘八	徳川家光	4	632	4
LBg3_00022	高野　孟	世紀末地球市民革命	5	792	4

表 4.3 は、500 語当たりの個別調整頻度（小数点以下繰り上げ）における、頻度上位 10 文書で、書名は表 4.2 とほぼ同じになっている。図 4.24 は過大調整がないから正確なわけではなく、過大調整は相対頻度にした段階ですでに生じているのだが、基準値を小さくしたため、その過大調整が見えにくくなっているに過ぎない。このしわ寄せは頻度 1 に現れ、本来、頻度を区別して分けた方がよいものまでがすべて 1 にまとめられている。それよりは、図 4.25 のように統合形式平均で調整し、過大調整となっている文書を特定しやすくした方が有益な分析ができると思われる。

同様の観察を頻度順位 500 位の「社会」でも行ってみよう。「そこ」と同じように、図 4.27 の固定長の文書度数分布と図 4.28 の基準値 500 語・四捨五入のグラフは、ほぼ同じ形状をしており、固定長と統合形式の比較には向くが、統合形式を使用したことによる情報量の増加が生かされていない。図 4.29 の基準値 500 語・小数点以下切り上げのグラフは、情報の切り捨てはないが、頻度 1 が非常に多く、きめ細かく分析できない。図 4.30 の基準値を統合形式平均にしたグラフは、過大調整された文書が現れるが、図 4.12 の散布図を見ても 2,000 語～3,000 語の区間で頻度 25 前後の文書が出現しているた

め、それほど多くの文書が実態と食い違っているわけではない。一方、図 4.31 のように基準値を 5,000 語にしたグラフは過大調整された文書が多い。これらの傾向は「そこ」で観察された傾向と同じである。

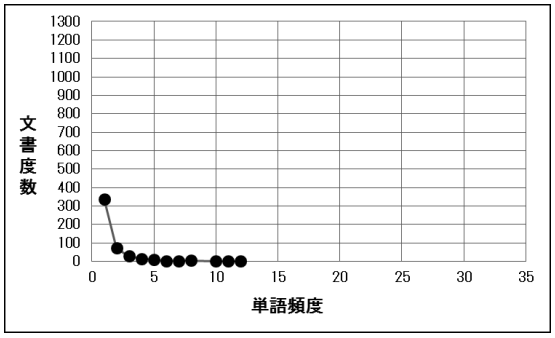


図 4.27 「社会」固定長

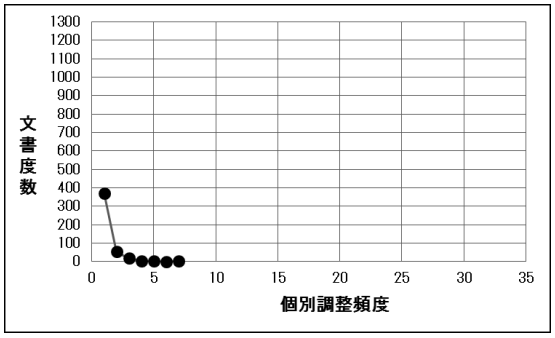


図 4.28 「社会」統合形式：500 語・四捨五入

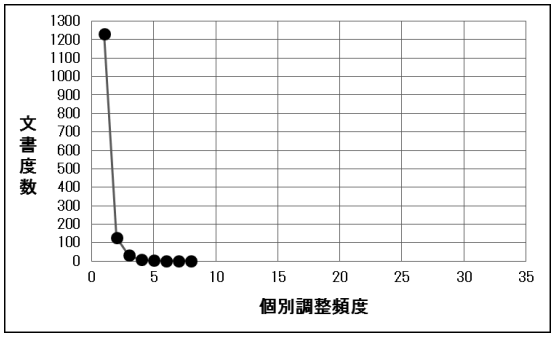


図 4.29 「社会」: 500 語・繰り上げ

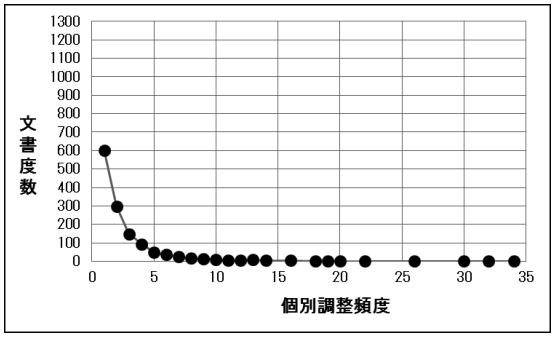


図 4.30 「社会」統合形式：平均・四捨五入

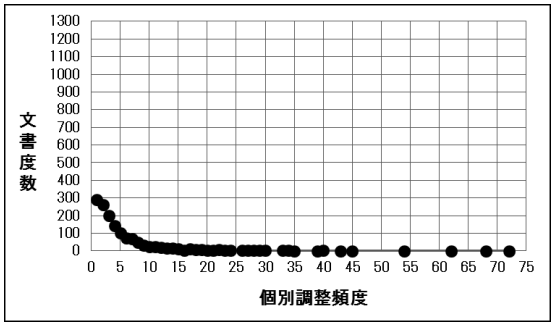


図 4.31 「社会」統合形式：5,000 語・四捨五入

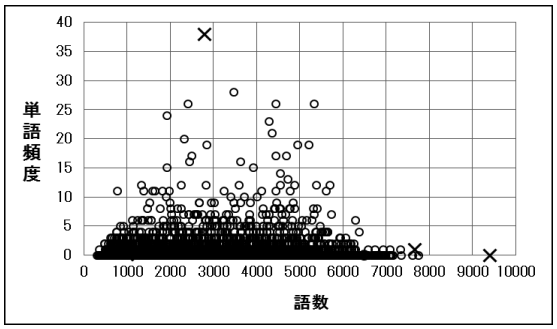


図 4.12「社会」の統合形式の散布図（再掲）

最後に頻度順位 3,010 位の「興味深い」でも同様の観察を行う。固定長の図 4.32 と統合形式を 500 語で調整して四捨五入した図 4.33 では、単語頻度が 1 しかないのは同じだが、図 4.33 では小数点以下を四捨五入したため、文書度数が固定長の半分になっている。ここでは、多くの情報が切り捨てられ、図 4.33 の分布図は、図 4.32 の固定長よ

り精度が低い分布図になっている。基準値 500 語・小数点以下切り上げの図 4.34 では、これとは逆に頻度 1 が非常に多く、情報の切り捨てはなくなったが、やはりきめ細かな分析が難しい。

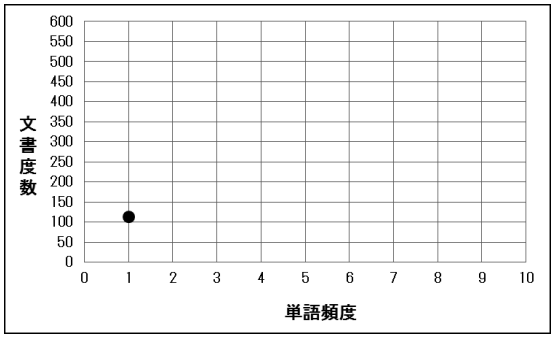


図 4.32「興味深い」固定長

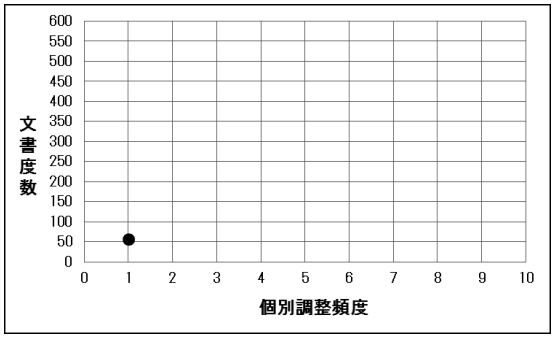


図 4.33「興味深い」統合形式
：500 語・四捨五入

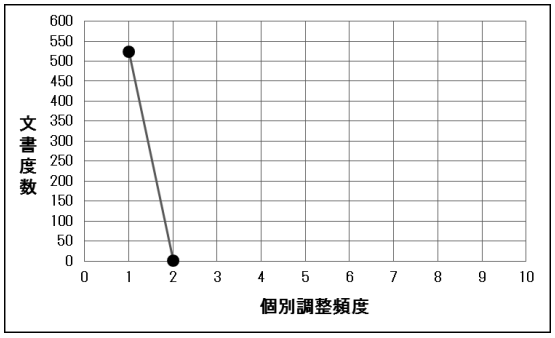


図 4.34「興味深い」統合形式
：500 語・繰り上げ

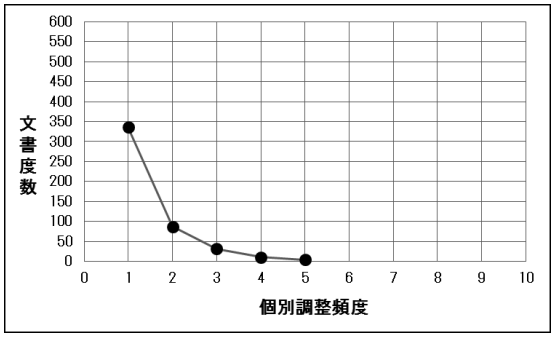


図 4.35「興味深い」統合形式
：平均・四捨五入

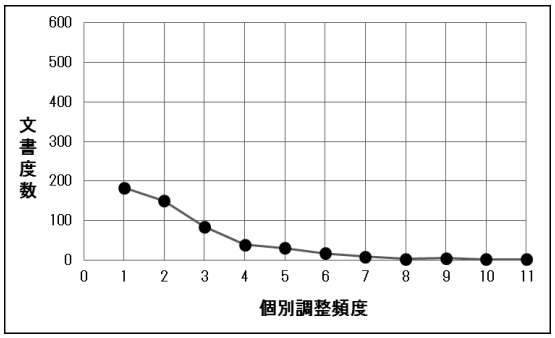


図 4.36 「興味深い」統合形式
：5,000 語・四捨五入

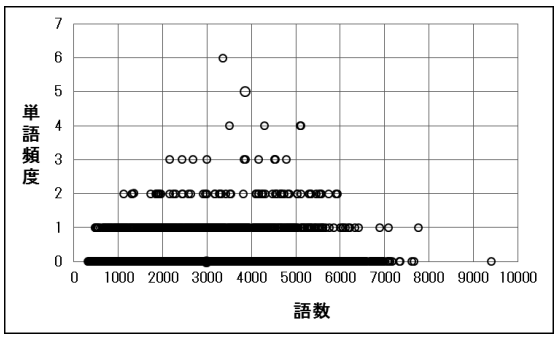


図 4.15 「興味深い」の統合形式の
散布図（再掲）

図 4.33 と図 4.34 を比較すると、頻度 1 の文書の差が 468 文書ある。これは個別調整頻度で 0.5 未満の文書が 468 文書存在したことを意味する。「興味深い」という単語は、文書の語数を多くすると出現する確率は増えるが、語数を多くしたからといって何回も

使用される単語ではない。このような単語の場合、図 4.32 の固定長の分布図自体が母集団の分布を正確に反映している可能性が低いため、固定長と統合形式の分布を比較すること自体が、有益な分析となりにくい。図 4.35 の基準値を統合形式平均にしたグラフは、頻度 4 以上の文書は過大調整された文書と思われるものの、図 4.15 の散布図から推測される調整頻度の状況に近い分布になっており、実態からのずれは少ない。図 4.36 のように基準値を 5,000 語にしたグラフは過大調整された文書が多く、頻度 2 の文書が頻度 1 の文書とそれほど変わらず出現するなど、図 4.15 の散布図から推測される調整頻度の状況とは異なっている。これらの傾向は、概ね「そこ」や「社会」と同じであるが、頻度順位 3,010 位、固定長頻度 114 の「興味深い」でも、固定長分布の代表性が低いと考えられる点には注意が必要である。

以上、個別調整頻度の算出方法と、それを使用して文書度数折れ線を描く方法を検討した。単語の多くは語数に連動して増減しないため、単語の頻度を相対頻度に直して分析すること自体、どこまで有効な分析になっているか不明である。しかし、語数の異なるコーパスを使用して分析するとしたら、現在のところ相対頻度を使用する以外に有効な方法がないのが現状である。先行研究で使用されてきた調整頻度は相対頻度に一定の基準値をかけた頻度で、何らかのカテゴリーごとに算出されることが多かった。本研究では個々の文書を調整した頻度を個別調整頻度と呼び、これを使用して文書度数分布を観察する方法を検討した。この個別調整頻度は、調整を行うための基準値をいくつに設定するか、小数点以下を四捨五入するか、繰り上げるかで文書度数折れ線の形状が大きく変化する。本研究の検討では、単独の文書度数折れ線を観察する場合であれば、コーパスの語数平均で四捨五入した値を使用して作図する方法が最も妥当だと考えられた。ただし、この方法では、頻度が高く調整された文書に過大調整が起きている可能性が高いため、これらを調査し、場合によっては分析に含めないなどの対策を施す必要がある。また、個別調整頻度を使用した文書度数折れ線は、粗頻度を相対頻度に直した段階で誤差を含み、さらに小数点以下の値を整数に直す段階でも誤差が生じるため、それほど正確な分布図になっていないことを念頭において分析に使用するべきである。

第 2 節で検討した語数と単語頻度の散布図であれば、このような誤差は生じないが、比較したい単語の散布図同士を見比べても、その違いがはっきりとは分かりにくい場合がある。文書度数分布の場合、比較したい対象の文書度数折れ線を一つの図にまとめて描くことも可能である。このため、散布図よりも文書度数折れ線の方が有効な分析ができる場合がある。文書度数折れ線を使用するか、散布図を使用するかは、分析目的によ

って使い分けることで、より有効な分析につながる。

第 5 節　必要文書数の見積もり

本節では、調査対象に対してどれぐらいの文書数があれば、母平均の正確な推定が可能になるのかという必要文書数の見積もりを行い、コーパスを使用して有意義な分析ができる調査対象の頻度の目安について考察する。第 3 節では、文書内の単語分布を観察し、語数の多い文書から語数の少ない文書を抽出した場合、それが元の文書における単語分布の正確な縮図となっている可能性はそれほど高くないことを確認した。このため、本研究では、語数が少ない固定長より、語数が多い統合形式の方がより正確な標本になっていると考える。しかし、統合形式は文書ごとの語数がばらばらであるため、これを平準化しないと分析に使いづらい。このため、前節では文書ごとの頻度を個別調整頻度によって平準化し、これを使用して文書度数分布を観察する方法を検討した。しかし、この方法によっても、多くの誤差が生じる可能性が示唆された。

ただし、第 3 節で述べたように、これらの誤差は、標本数が多くなれば互いに打ち消し合い、大数の法則によって標本平均は母平均に近づくと考えられる。そこで本節では、そもそもどれぐらいの標本数があれば、大数の法則が働いて正確な値となるのかを見積もり、コーパスにどれぐらい出現する単語であれば、有意義な分析ができるのかという目安について考察する。

Biber（1993）では、コーパスに出現した調査対象の 1 文書当たりの平均頻度と標準偏差を使用して、誤差 5%以内で母集団の頻度が推定できる文書数が見積もられている。Biber（1993）による必要文書数の推定方法は、平均 μ、分散 σ^2 に従う母集団からサンプルサイズ n の標本を抽出する場合、その平均値 $\bar{x}$ の分布は n が大きくなるにつれて正規分布 $N(\mu, \sigma^2/n)$ に近づくという中心極限定理に基づいている。Biber（1993）に基づき、s を標準偏差、$\bar{x}$ を標本平均として書き直した式が（式 4.2）である。

（式 4.2）
$$n = s^2 / (0.025\,\bar{x})^2$$

（式 4.2）では、分子の s（標準偏差）、すなわち文書間で出現する頻度のばらつきが大きければ大きいほど多くの文書が必要になり、また分母の $\bar{x}$、すなわちコーパスに出現した頻度の平均が小さければ小さいほど多くの文書が必要になる。つまり（式 4.2）は、低頻度で文書間の頻度にばらつきがある調査対象ほど、正確な推定を行うためには

多くの文書を必要とすることを意味している。

表 4.4 は、5%の誤差で 7 種類の調査対象を分析するために必要な文書数を推定した Biber（1993:254）の分析結果である。表中の Tolerable error は、平均に誤差の 5%をかけた値である。ここでは文書の語数が 1,000 語、文書数が 481 のコーパスに出現した頻度平均と標準偏差が使用されている。これを見ると、1 文書当たり平均 180.5 語が出現する Nouns（名詞）では、わずか 59.8 文書あれば誤差 5%以内の推定をすることが可能である。一方、Conditional clauses（条件節）のように 1 文書当たり平均 2.5 回しか使用されない調査対象では、1,190 文書が必要となる。Brown コーパスの文書数は 500 であるから、これらのコーパスでは条件節の分析は 5%以内の誤差の範囲では分析が困難だと考えられる。

表 4.4　Estimates of required sample sizes (number of texts) for the total corpus

Biber（1993:254）より引用

	Mean score in pilot corpus	Standard deviation in pilot corpus	Tolerable error	Required *N*
Nouns	180.5	35.6	9.03	59.8
Prepositions	110.5	25.4	5.53	81.2
Present tense	77.7	34.3	3.89	299.4
Past tense	40.1	30.4	2.01	883.1
Passives	9.6	6.6	0.48	726.3
WH relative clauses	3.5	1.9	0.18	452.8
Conditional clauses	2.5	2.2	0.13	1190.0

表 4.5 は、BCCWJ 図書館 SC・固定長・長単位を使用して、8 種類の単語に対して同様の分析を行った結果である。表 4.4 に加えて、頻度順位、頻度、出現した文書数、誤差 5%の分析に必要な頻度を書き加えた。なお図書館 SC の 1 文書当たりの長単位の語数は平均 523.6 語、文書数は 10,551 である。

図書館 SC の文書数は 10,551 だが、順位 500 位の「社会」の必要文書数は 60,909 文書となっているため、BCCWJ ではこれほどの高順位の単語でも 5%の誤差の範囲で分析することが困難であることが分かる。「乗り物」や「虫歯」などはごく身近な単語だが、コーパスでの出現数は少なく、80 万～200 万の文書が必要になる。「喫緊」に至っては、16,880,000 文書が必要である。「喫緊」の頻度に対して 5%の誤差の範囲で分析することができるコーパスを代表性を持ったコーパスと考えるなら、1 文書当たりの語数

が 500 語の場合、約 1,700 万文書が必要であり、語数にすれば 84 億語のコーパスが必要だという計算になる。

表 4.5　BCCWJ 図書館 SC・固定長の出現数をもとにした
誤差 5%の分析に必要な文書数の推定[8]

単語	順位	頻度	文書数	平均	標準偏差	許容可能誤差	必要文書数	必要頻度
の	1	340,436	10,550	32.266	10.338	1.613	164	5,299
そこ(代名詞)	100	4,289	3,143	0.407	0.732	0.020	5,190	2,110
社会	500	762	472	0.072	0.446	0.004	60,909	4,399
戦う	998	388	307	0.037	0.236	1.8E-03	66,131	2,432
興味深い	3,010	114	114	0.011	0.103	5.4E-04	146,484	1,583
乗り物	9,890	28	24	0.003	0.058	1.3E-04	773,576	2,053
虫歯	20,050	12	10	0.001	0.041	5.7E-05	2,108,600	2,398
喫緊	113,318	1	1	9.5E-05	0.010	4.7E-06	16,880,000	1,600

これと同様の方法を統合形式・長単位を使用して行ったのが表 4.6 である。文書の値は統合形式平均の 2378 語で平準化した個別調整頻度を使用した。小数点以下の値は四捨五入せず、そのまま使用している。

表 4.6 BCCWJ 図書館 SC・統合形式の出現数をもとにした
誤差 5%の分析に必要な文書数の推定

単語	順位	頻度	文書数	平均	標準偏差	許容可能誤差	必要文書数	必要頻度
の	1	1,543,392	10,550	146.29	40.947	7.315	125	18,337
そこ(代名詞)	95	20,165	6,865	1.846	2.191	0.092	2,255	4,163
社会	495	3,466	1,405	0.333	1.392	0.017	27,977	9,315
戦う	920	1,886	1,111	0.175	0.721	8.8E-03	27,057	4,746
興味深い	4,071	629	527	0.061	0.325	3.0E-03	45,773	2,778
乗り物	7,567	174	113	0.016	0.207	7.8E-04	284,118	4,417
虫歯	19,595	56	37	0.006	0.140	2.8E-04	1,007,211	5,602
喫緊	157,200	3	3	5.0E-04	0.039	2.5E-05	9,833,300	4,899

表 4.6 を見ると、多くの単語で必要文書数が半分程度に下がっているのが分かる。これは、出現する頻度が増え、ばらつきも少なくなるからである。しかし、統合形式を使用しても、「社会」や「戦う」の必要文書数は 27,000 文書を超え、図書館 SC では誤差 5%以内の分析が難しい。このような結果は衝撃的ではあるが、第 3 節で行った文書内の単語分布の観察内容からすると、必ずしも納得できない結果ではない。

[8] 国立国語研究所で編纂した長単位語彙表データ（頻度 2 以上）：BCCWJ_frequencylist_luw2_ver1_0.zip（http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html）では、「興味深い」は頻度 64、順位 4979 位となっているが、検索の結果、頻度 114 であったため、順位を修正した。

第 1 章では、「喫緊」が新聞と白書のどちらの媒体に出現しやすいかという後藤（2007）の追試を行った。第 1 章の調査では、（新聞，白書）の順に示した固定長の調整頻度で（3.22，2.88）、統合形式の調整頻度で（2.19，3.69）と、使用するデータによって結果が逆になった。本章で扱った個別調整頻度を使用し、500 語当たりの文書平均を求めると、平均値は同様に、固定長で（0.00158，0.00146）、統合形式で（0.00107，0.00159）となり、類似の結果になる。本章の考察からすれば、固定長より統合形式の結果の方が正確だと思われるが、これほど低頻度の単語の場合、この結果に意味があるとは考えにくい。

コーパス研究では、初期の頃から、高頻度語であれば小さなコーパスでも概ね正確な値になるといわれてきた。本節の分析結果からすると、語数約 3,000 万語の BCCWJ 図書館 SC・統合形式でも、高頻度語以外を分析するためには十分なサイズにはなっていないと考えられる。どれぐらいの頻度順位の単語であれば、誤差 5%以内の分析が可能であるかは、その単語の平均値と標準偏差に影響されるため、一概に示すことができないが、たとえば補助動詞「てある」の場合、複合形式の頻度順位が 353 位、合計頻度が 4,827 語、出現文書が 2,782、平均 0.64 語、標準偏差 1.55 語で、これを基に推計した必要文書数は 9,511 となる。図書館 SC の文書数は 10,551 であるから、「てある」であれば、誤差 5%で推定できる。頻度順位 352 位、合計頻度 4,833 語の副詞「やはり」だと、出現文書が 2,953、平均 0.63 語、標準偏差 1.41 語で、必要文書数が 7,923 である。平均語数は似ているが、「やはり」は、文書ごとのばらつきが少ない分、必要文書数が少なくなっている。一方、頻度順位 362 位、合計頻度 4,632 語の名詞「名前」の場合、出現文書が 2,308、平均 0.62 語と、平均値は「てある」や「やはり」に近いものの、標準偏差が 2.36 語とばらつきが大きく、必要文書数が 23,071 となっている。

第 2 章でも述べたように、言語研究で統計分析を行う目的は、正確な頻度を推定することだけでなく、その調査対象がどのような種類の文書に出現しやすいのかといった分布の偏りを観察することにも求められるため、必ずしも誤差 5%の精度が必要とされるわけではない。しかし、調査対象の頻度が数百語以下の場合、それほど有益な分析にはならない可能性を考えておく必要がある。

BCCWJ には図書館 SC のほかに、国立国会図書館の 5 年分の蔵書を母集団にした出版書籍が存在する。この二つは、書籍の流通実態と生産実態の解明のために、母集団を別にして設計されたが、集積されている文書自体はよく似た傾向を持っている。出版書籍のサンプルサイズは 10,117 文書であるため、図書館 SC で必要文書数が不足する調査

対象については、筆者が森（2016）でその可能性を示したように、この二つのレジスターを組み合わせて使用することも、検討してみる価値が高い。少なくとも、互いの分布を比較すれば、おおよその誤差の程度を知る手がかりが得られる。

第 6 節　まとめ

本章では調査対象の文書分布を観察することにより、データクリーニングが必要な外れ値を発見するとともに、調査対象の言語的特徴を把握する方法を検討した。言語単位を個体と考える従来の分析法では、個体の観測値は全て 1 であるため、分布が生じない。しかし、文書を個体と考える本研究では、個体はさまざまな観測値を取るため、その分布を観察することで有益な情報が得られる。

固定長を使用した文書度数分布では、固定長に異常に語数が多い文書が混じっている実態や、表の中身をデータに含めたため、「色素」のように大量の単語が一つの文書に出現する例が観察された。また、「の」は正規分布に近い分布を示すのに、「社会」などの話題語は L 字型の分布となるなどの言語的特徴が観察された。統合形式を使用した語数と単語頻度の散布図では、外れ値となっている文書の内容を確認することで、本来は接続詞に形態素解析されるべき「そこで」が、代名詞＋格助詞に解析されている実態や、固定長では 1 文書当たり 1 回の使用に限られた「興味深い」が、実は連続して使用されている例などが発見できた。このように、文書の分布を観察することで、分析を開始する前のデータクリーニングや、言語的特徴の把握が行える。

第 1 章では、「言語研究者が，読んだこともないテクストを研究することほど，矛盾に満ちたものはない」という伊藤（2005:96）の指摘を引用した。大量に言語データを集積したコーパスにおいて、全ての文書を読むことはほぼ不可能に近いが、本章で示したように、文書の分布図を描いて観察すれば、特徴的な文書に的を絞ってテキストを読むことが可能になる。つまり、言語単位を個体と考えて分析していた段階ではブラックボックスだったコーパスの中身を、効果的・効率的に確認しながら分析していくことが可能になる。

外側から見ていた段階では、巨大なデータを集積した正確なデータベースのように感じられたコーパスも、その内部に分け入って実際に集積されている文書を観察してみると、中には異常と思われるデータも存在しており、サンプリングの誤差も想像以上に大きい可能性が示唆される。語数の異なるデータを比較する際に必須の統計量である調整頻度にしても、どこまで正確な調整ができるのか、必ずしも明確ではない。コーパスの

分析では、このような誤差や曖昧さが内包されていることを前提とし、調査対象をできるだけ高頻度語に絞る方が安全だと考えられる。頻度が低い言語現象については、性急に結論を出すのではなく、その現象が十分に調査できるコーパスの出現を待つことも必要であろう。また、筆者が森（2016）で検証を試みた BCCWJ の図書館 SC と出版書籍の混合分析などをさらに精密化させ、少しでも文書数を増やすような分析法を検討していくことが望まれる。

第 5 章　代表値と分布図を併用した頻度比較の方法

本章では、代表値を使用した分析の方法を検討し、代表値単独ではなく、分布図と併用して分析する方法を提案する。第 4 章の分布観察では、文字数がほぼ一定に揃っている固定長より、文字数が不揃いな統合形式を使用して分析する方が困難であった。このことは、代表値を使用した分析でも、同様だと考えられる。その一方で、現在製作されているコーパスは、文書の文字数が一定に揃えられている均衡コーパスより、文字数が不揃いな広義コーパスの方が圧倒的に多い。また、広義コーパスでの分析法が明らかになれば、均衡コーパスの分析はそれに準じて行うことができる。そこで、本章では、広義コーパスに属する学習者コーパスを対象に、代表値を使用した分析の方法を検討する。

学習者コーパスを使用した主要な研究の一つに、学習者のレベルごとの習得状況を明らかにする分析がある。本章では条件表現の「たら」が、どのレベルでどれぐらい習得されているかを比較する際、①調整頻度、②平均値、③中央値、④「たら」を使用した学習者数という 4 種類の代表値を使用すると、どれほど有効な比較が行えるのかを検証する。また、代表値を使用した分析に加え、作図による分布観察の方法を検討し、これらを併用した分析法の提案を行う。

第 1 節では、日本語教育においてこれまで最も多く使用されてきた KY コーパスを使用して分析を行う。第 2 節では今後の日本語教育において多用されることが見込まれる I-JAS を使用して同様の分析を行い、分析の一般化を行う。第 3 節でまとめを述べる。

第 1 節　頻度分析法の比較：KY コーパスの場合

はじめに KY コーパスを使用し、条件表現「たら」の習得レベル別の頻度比較を行う。第 1.1 項では分析に使用するデータに関する説明を行う。第 1.2 項では①調整頻度、②平均値、③中央値、④「たら」を使用した学習者数、およびそれらの割合を算出し、分析結果を比較する。第 1.3 項ではこれらの代表値の妥当性を検討する。第 1.4 項では代表値と分布図を併用した分析法の提案を行う。

第 1.1 項　使用するデータの説明

KY コーパスは、鎌田修・山内博之の両氏によって構築された学習者コーパスである。データには、OPI（Oral Proficiency Interview）のスクリプトが使用され、1999 年に一般公開された（鎌田，1999）。語数はおよそ 17 万語、学習者の内訳は初級下～超級までの 9 レベルに分けられた英語・韓国語・中国語母語話者各 30 名の合計 90 名となっている。

2008 年には李在鎬氏等によって形態素解析と誤用タグが施されたデータが公開され（李，2009）、2013 年に検索システムを備えた「タグ付き KY コーパス」として Web 公開されている[9]。KY コーパスの内容や分析上の注意点については、鎌田（1999，2006）、山内（1999）に詳しい。以下、これらに基づいて概要を記す。

KY コーパスは製作者を含めて 25 名のテスターから提供を受けた OPI データに基づいて構築されている。OPI は学習者に対して最長 30 分のインタビューを行い、米国外国語教育協会（ACTFL）が定めた外国語能力基準によって学習者の能力を判定する評価法である。学習者のレベルは KY コーパス製作当時の基準で超級、上級（上級上・上級）、中級（上・中・下）、初級（上・中・下）の 9 段階[10]になっており、これらの総合判定規定は以下のとおりである。

> ● 超級（Superior）：意見の裏付け，仮説構築，具体的・抽象的話題について議論ができ，そして言語的に不慣れな状況が処理できる
> ● 上級（Advanced）：すべての時間的枠組で叙述，描写ができ，かつ，複雑な状況が処理できる
> ● 中級（Intermediate）：自分なりに言葉が使え，なじみ深い話題について簡単な質問をしたり，答えたりでき，また簡単な状況や，やり取りに対処できる
> ● 初級（Novice）：決まり文句や習い覚えた語句，単語の羅列で最小限のコミュニケーションが行える （鎌田，2006:47）

テスターは学習者がこれらのレベルのどれに該当するのかを判定するため、学習者が発話し続けるのが困難なレベルにまで会話を誘導する「突き上げ」を行って学習者の言語的挫折を引き出す。またテスト中に学習者が中級以上の能力を持つと判断された場合には、ロールプレイが行われる。このためインタビューはできるだけ自然な会話を目指しながらも、通常の雑談とは異なった内容を含むことになる。このため分析に当たっては注意が必要であるという（鎌田，1999:234-6）。以下、その主なものについて概要を記す。

・OPI は学習者能力の上限と下限を判定するため、データに言語的挫折が生じている部

[9] http://jhlee.sakura.ne.jp/kyc/

[10] 1999 年に改訂され、上級も上・中・下の 3 段階になったため、現在は 10 段階となっている。

分と問題なく話し続けられる部分の両方を含むが、その分量は学習者やテスターによって異なるため、誤用・正用の割合を学習者間で比較することはあまり意味がない。

・各レベルに必要なタスクの種類が決まっているため、たとえば上級では「記述・叙述・意見」を求めるタスクに必要な「～と思う」などの形式は頻出するが、伝達文の「～そうだ」などの形式は自然発生的にしか出てこない。

・それぞれの OPI はそれぞれ独自の話題で展開していくため、話題に左右される語彙の使用頻度はそれぞれのデータによって大きく異なり、一般化が困難である。

以上の設計内容からすると、KY コーパスは母集団を定めてそこから無作為抽出されたデータではなく、また、学習者の発話を引き出す方法も OPI の実施方法によって一定の偏りを持っていることが分かる。

OPI のデータを使用する最大の利点は、学習者のレベルが ACTFL の基準によって判定されていることにあるが、これについてもテスターによる偏りが考えられる。山内（1999:244）ではこの点について次のように述べられている。

> 90 本という、かなり多くのテープを集めたため、それぞれのすべてが、非のうちどころのない OPI インタビューであるというわけではない。インタビューの行い方が稚拙なものもあれば、また、判定結果の信頼性にやや疑問が持たれるようなものも、若干はある。（山内，1999:244）

KY コーパスを使用して分析を行う場合は、このような問題点を踏まえて、分析していく必要がある。

次に KY コーパスの量的な側面を検討する。1999 年公開の KY コーパス Ver.1.1 は、形態素解析を行っていないプレーンテキストのデータであるため、当時はデータ量がどれぐらいあるか確定できていなかった。鎌田（2006:43）では OPI のインタビューにかかる時間の大半が 20～30 分で、総数 90 名分であるところから、中間の 25 分×90 名で概算し、「KY コーパスは総時間 2250 分ほどの音声データを文字化したものといえる」と述べられている。

KY コーパスの形態素解析は複数の研究者によって試みられているが、2008 年に李在鎬氏等によって形態素解析された研究が最も精度が高いと考えられる（山内, 2015:50）。

李在鎬氏等の研究に基づいて報告されたデータ量は、李・淺尾・濱野・佐野ほか（2008）では 173,198 形態素、李（2009）では 232,605 語と記されている。筆者が「タグ付き KY コーパス」を使用して学習者の全品詞をダウンロードして集計した形態素数は 170,454 であった[11]。

表 5.1 は筆者がダウンロードした品詞数のうち、記号を除いた語数を使用して算出した形態素数の統計量である。これを見ると KY コーパスではレベル別の人数や語数にかなりのばらつきがあることが分かる。レベルの大分類では初級 5、中級 10、上級 10、超級 5 のようにある程度人数が揃えられているが、下位分類ではばらつきがある。また学習者の語数も大きく異なっている。Brown コーパスや BCCWJ 固定長のような均衡コーパスでは、データの語数や文字数が一定に揃えられているため、統計的な分析に適しているが、学習者コーパスは BCCWJ でいえば可変長や統合形式に当たり、データの長さがさまざまに異なっているため統計分析が難しい。

表 5.1　KY コーパスの母語とレベル別の人数・語数平均・標準偏差

	英語			韓国語			中国語		
レベル	人数	語数平均	標準偏差	人数	語数平均	標準偏差	人数	語数平均	標準偏差
初級下	1	108	−	2	189	93	1	207	−
初級中	2	763	615	1	513	−	2	252	104
初級上	2	656	465	2	622	181	2	705	121
中級下	4	1,122	326	2	1,183	477	3	1,495	665
中級中	4	1,509	556	6	1,477	297	4	1,490	333
中級上	2	1,908	630	2	1,890	700	3	2,136	421
上級	3	1,755	414	6	1,921	303	3	2,493	435
上級上	7	2,808	610	4	2,671	592	7	2,697	584
超級	5	2,757	808	5	3,257	812	5	2,584	370

特に初級は学習者数が少なく、その語数も非常に少ない点に注意が必要である。初級は「決まり文句や習い覚えた語句, 単語の羅列で最小限のコミュニケーションが行える」（鎌田, 2006:47）というレベルであるから、発話量（語数）がごく少ない。OPI の場合、レベル判定が確定すればそこでインタビューは終わり、制限時間の 30 分間まで会話をし続けることはない。このため習得レベルの低い学習者に語数の少ないデータが多いと考えられる。

本章で代表値の検討に用いるデータは、Web 上で検索が行えるタグ付き KY コーパスを利用して取得した。条件表現「たら・だら」の用例を得るために、文字列「たら・だ

[11] 2017.04.28 閲覧。

ら」を検索し、目視で調査対象外用例と誤用を除いた結果、580 の用例が得られた。以後の分析にはこの 580 のデータを用いる。

学習者コーパスを使用した計量的な分析では、グレンジャー等によって提唱されている対照中間言語分析（contrastive interlanguage analysis：CIA）が主流で、学習者コーパスと母語話者コーパスを比較して学習者の過剰・過少使用を探る分析法が一般的である（グレンジャー（編），2008:14-6；石川，2017:67-8）。しかし KY コーパスには母語話者のデータが付随していないため、本章では名大会話コーパスを使用して母語話者のデータを取得した[12]。

名大会話コーパスは科学研究費基盤研究（B）（2）「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（平成 13 年度～15 年度　研究代表者　大曽美恵子）の一環として製作された会話コーパスで、10 代から 90 代までの日本人母語話者（女性 161 名、男性 37 名）による雑談 129 会話、約 142 万語のデータである。現在は国立国語研究所の検索インタフェース「中納言」による検索が可能になっている。本章ではこれを利用し、語彙素「た」、品詞「助動詞」、活用形「仮定形」で検索した「たら」4,491 語のデータを使用する。

名大会話コーパスは一人の話者が複数の会話に登場する場合もあれば、一つの会話で短い発言しかしない場合もある。データは一つの会話が 30 分～1 時間程度になるように録音されたもので、データを収集した観察単位は話者ではなく会話になっている。そこでデータは会話単位で集約し、1 会話当たりの平均値や中央値などを算出した。KY コーパスと名大会話コーパスを比較する意味は、母語話者がさまざまな雑談を行った場合に平均的に使用する「たら」の頻度に比べ、学習者のレベルによってどれぐらいの過少・過剰使用があるかを評価することにある。ただし KY コーパスは純粋な雑談ではなく、面接官によって管理されたインタビューであるため、厳密な比較は困難で、あくまで参考程度の比較にとどまる。

また、タグ付き KY コーパスは形態素解析エンジン ChaSen（茶筅）、辞書 IPAdic で解析されている（李、2009:62）。一方、名大会話コーパスや I-JAS は形態素解析エンジン MeCab、辞書 UniDic で解析されている[13]。この二つの辞書では言語単位の認定方法が異

[12] 2017.04.28 閲覧。

[13] 国立国語研究所の「中納言オンラインマニュアル」に，「中納言に格納された短単位データの作成は自動形態素解析によって行われています。 形態素解析処理は形態素解析器に「MeCab」、解析用辞書に「UniDic」を使用しています」とある。（http://pj.ninjal.ac.jp/corpus_center/chu-00.html，2017.11.19 閲覧）。

なり、たとえば「経済学部」という語なら、IPAdic は 1 単位、UniDic では「経済／学／部」の 3 単位で解析される（小木曽、2014:100）。また単位名も IPAdic では形態素、UniDic では短単位[14]という名称になる。このため、これらも厳密な比較は困難で、データ数の名称もそれぞれ形態素数、短単位数と呼ぶのが正確だが、本研究ではこれらを単純化して語数と呼ぶ。

第 1.2 項　代表値を使用した頻度比較の結果

はじめにタグ付き KY コーパスと名大会話コーパスの基礎統計量と、「たら」の調査結果に基づいて算出した①調整頻度、②平均値（個別平均）、③中央値、④「たら」を使用した学習者数（使用者数）という 4 種類の代表値、およびその割合①’～④’の分析結果を表 5.2 に示す。

表 5.2　タグ付き KY コーパス・名大会話コーパスの基礎統計量と代表値の分析結果の比較

				⓪	①	②	③	④	①’	②’	③’	④’
レベル	学習者数	語数合計	語数平均	タラ粗頻度	調整頻度	個別平均	中央値	使用者数	調整頻度割合	個別平均割合	中央値割合	使用者割合
初級下	4	692	173.0	0	0.0	0.0	0.0	0	0%	0%	0%	0%
初級中	5	2,542	508.4	0	0.0	0.0	0.0	0	0%	0%	0%	0%
初級上	6	3,965	660.8	0	0.0	0.0	0.0	0	0%	0%	0%	0%
中級下	9	11,338	1,259.8	13	1.1	0.8	0.0	3	33%	24%	0%	33%
中級中	14	20,854	1,489.6	42	2.0	2.0	1.6	13	57%	56%	50%	93%
中級上	7	14,004	2,000.6	76	5.4	5.2	5.0	7	154%	150%	156%	100%
上級	12	24,269	2,022.4	80	3.3	3.3	2.3	12	94%	96%	70%	100%
上級上	18	49,213	2,734.1	186	3.8	3.6	3.4	18	108%	103%	106%	100%
超級	15	42,990	2,866.0	183	4.3	4.2	3.4	15	121%	121%	106%	100%
母語	＊129	1,419,729	11,005.7	4991	3.5	3.5	3.2	129	100%	100%	100%	100%

＊ 母語の 129 は人数ではなく名大会話コーパスの会話数。この行は 1 会話当たりの統計量を算出している。

KY コーパスのレベル別学習者数は、先に述べたように一定数にはなっていない。語数合計はレベル別の語数合計で、初級下では 692 語、語数平均はこれを 4 名で割った値の 173.0 語である。

⓪粗頻度は、学習者の「たら」の用例から調査対象外用例と誤用を除く以外、何も調整していない頻度である。粗頻度で見ると中級上の粗頻度は 76、上級の粗頻度は 80 であるから、一見中級上より上級の方が「たら」を多く使用しているように見える。しか

[14] 言語研究には短単位より文節に基づいた長単位の方が適しているとされるが（国立国語研究所コーパス開発センター，2015:26；小椋秀樹，2014:73-8）、I-JAS や名大会話コーパスには長単位データが用意されていないため、短単位を使用して分析を行う。

し中級上の学習者は 7 名、上級の学習者は 12 名であるから、一人当たりの「たら」の頻度で考えると中級上の方が多い。また語数合計で考えると中級上は 14,004 語発話した中で 76 回「たら」を使用しているのに対し、上級は 24,269 語発話した中で 80 回「たら」を使用しているのであるから、両者の語数を調整して比較すれば、やはり中級上の方が多くなる。つまり、レベルごとに人数や語数に違いがある場合、粗頻度で比較しても有効な比較は困難であり、レベルごとの条件を一定にした上で比較する必要がある。表 5.2 の①調整頻度は「粗頻度合計÷語数合計×1,000 語」で平準化した 1,000 語当たりの調整頻度である[15]。これを見ると中級上では 5.4、上級では 3.3 で、上級より中級上の方が 1.63 倍ほど多く使用されている。

②個別平均（個別調整頻度の平均値）は学習者一人一人の個別調整頻度を出し、その合計を人数で割って一人当たりの平均を求めた値である。調整頻度と個別平均はよく似た値になるが、中級下では 1.1 と 0.8、中級上では 5.4 と 5.2 のように若干異なっている。この理由は次項で詳しく述べるが、調整頻度が合計数を使用して簡便に計算しているため、正確な値にならないことに起因している。

③中央値は、学習者ごとの「たら」の個別調整頻度を大きさの順に並べた時、真ん中に位置する値である。学習者が偶数の場合は真ん中の 2 名の平均である。データの中に少数の極端な値（外れ値）が存在する場合は、平均値より中央値の方が代表値としてふさわしいとされている（吉田，2001:45）。

④使用者数は「たら」を 1 回でも使用した学習者が何名いるかを数えた値である。初級下、初級中、初級上の使用者は 0 名、中級下では 9 名のうち 3 名が「たら」を使用している。中級上からは全員が「たら」を使用している。このように各レベルの学習者全員のうち、何割の学習者が「たら」を使用しているかを計算した値が、表 5.2 の右端の④’使用者割合である。

①’調整頻度割合、②’個別平均割合、③’中央値割合は、各代表値における母語の値を 100%とした場合の各レベルの割合を示した値である。ここでは名大会話コーパスの値を基準にし、それと比較することで各レベルがどれぐらいの割合になっているかを

[15] 第 4 章では、個別調整頻度を使用して度数折れ線を描く場合は、データの平均値を使用するときめ細かな観察がしやすいこと述べた。これに従えば KY コーパスの語数平均である 1,887 語で調整することが考えられるが、代表値や散布図を使用する場合は、度数折れ線を描く場合と異なり、小数点以下を四捨五入する必要がないため、データの切り捨ては生じない。また、「たら」の場合、1 学習者当たりの出現数が多いため、1,000 語で調整しても分布の違いが明確に分かる。また、以下で代表値の比較を行う際、1,000 語で調整していると暗算しやすく、議論が理解されやすいため 1,000 語で調整している。

算出している。

調整頻度、平均値、中央値、使用者割合などの代表値は、どれも基本的な代表値であり、これまでそれぞれの代表値の妥当性が問題にされることは少なかった。しかし表 5.2 を見ると、どの代表値を使用するかで学習者が「たら」を習得した時期の判断が変わる可能性がある。たとえば①’調整頻度割合や②’個別平均割合なら、中級上から習得に至ったと判断できそうである。中級上の調整頻度割合は 154%、個別平均割合は 150%で、母語よりも多く「たら」が使用されている。上級ではわずかに割合が下がるが、ほぼ母語話者と同じレベルである。一方③’中央値割合はこの二つより判断が難しい。中級上では母語の 156%、上級では 70%と割合が大きく揺れている。中級上で過剰使用されているだけなら、習得したばかりで使用に偏りがあったとも考えられるが、上級で逆に過少使用になるのであれば、データの信憑性も含めて詳しく検討する必要がある。あえて③’の値だけで判断するなら、中央値では上級上と超級で母語話者と同レベルの 106%になるため、上級上で習得に至ったと考えるのが妥当かもしれない。④’使用者割合は中級中で 93%、中級上以上は 100%となっており、中級中から習得に至ったと判断してよさそうである。中級中では 1 名＝7%が「たら」を使用していないが、学習者コーパスの場合、未出現＝未習得とは必ずしも考えられない（石川、2012:220）。中級中では 1 名の学習者を除いて 93%の学習者が「たら」を使用していることから、このレベルでは「たら」の習得に至っていると判断してもよさそうである。

以上の例から分かることは、基本的な代表値であっても、どれを使用するかで結論が変わる可能性があるということである。分析者が無批判に何らかの代表値を使用した場合、その代表値が妥当な結論を導き出しているとは限らない。このため、それぞれの代表値の妥当性を検討してみる必要がある。

第 1.3 項　学習者の習得レベル別代表値の妥当性

表 5.2 の検討により、①～④の代表値ではどれを使用して分析するかで結論が変わる可能性があることが分かった。この四つの代表値の妥当性を評価するため、本項では表 5.2 の中級下の詳細データをまとめた表 5.3 を使用し、それぞれの代表値の計算方法を検討する。表 5.3 の 1 行目は学習者一人一人が発話した語数である。最も少ない学習者は 714 語、最大は 1,991 語で、同じ中級下でも 3 倍近くの幅がある。2 行目の粗頻度は学習者が「たら」を使用した回数で、何も調整していない頻度である。1 回も「たら」を使用していない学習者が 6 名いる。3 行目の個別調整頻度は、学習者個々に「たら」

の 1,000 語当たりの調整頻度を求めた値で、2÷1,520×1,000＝1.3 のように計算してある。

表 5.3 KY コーパス中級下の詳細データ：学習者 9 名の発話語数・「たら」の粗頻度・個別調整頻度

	中級下の学習者の個別データ									合計	人数	平均値	調整頻度
発話語数	714	739	846	1,018	1,299	1,455	1,520	1,756	1,991	11,338	9	1259.8	
粗頻度	0	0	0	0	0	0	2	9	2	13.0	9	1.4	1.1
個別調整頻度	0	0	0	0	0	0	1.3	5.1	1.0	7.4	9	0.8	

縦線より右側は、左側のデータに基づいて算出した値である。合計は 9 名の合計数、人数はいずれも 9 名、平均値は合計を人数で割った値である。右端の調整頻度は「たら」の粗頻度合計 13.0÷語数合計 11,338×1,000 で、1,000 語当たりの調整頻度である。これは粗頻度の平均値 1.4÷語数の平均値 1,259.8×1,000 と同じ値で、調整頻度とは、粗頻度平均を 1,000 語当たりの頻度に直した値と同じである。3 行目の個別調整頻度では、左側で一人一人の値を個別に調整しているため、右側で調整頻度を算出する必要はなく、合計を人数で割った値の 7.4÷9＝0.8 が個別調整頻度の平均値（個別平均）になる。

調整頻度も個別平均もどちらも平均値であるが、値が異なる（表 5.3 の網掛け箇所）。それは個別平均が個々の学習者ごとに調整頻度を算出した上で平均値を出しているのに対し、調整頻度は「たら」の粗頻度と学習者の語数を合計して割っているためである。表 5.3 の個別データの特徴は、語数の多い学習者だけが「たら」を使用しており、語数が少ない学習者は「たら」を使用していない点にある。この特徴の意味するところを、学習者の用例で具体的に観察してみよう。中級下で最も語数が少ないのは、ID が EIL01（E＝英語、IL＝中級下、01=№01）の学習者である。次の(1)はこの学習者のデータの一部で、T が面接官、S が学習者である。引用部では面接官が学習者に対し、学習者が現在住んでいる京都で面白いところはどこかを尋ねている。これに対する学習者の応答は、ほとんど単語でなされている。

(1) Ｔ：どんなところが、面白いですか、銀閣寺の面白いところ
Ｓ：どんな建物
Ｔ：建物、〈うん〉寺の形ね、のうち、〈うん〉でなにがあなたは面白いですか
Ｓ：おもしろい
Ｔ：きょうみ、興味がありますか

S：神社
T：神社で、あのー、銀閣寺、庭がいいですか、あのー
S：建物
T：建物がいいです
S：いいです　（KY コーパス，EIL01）

次の(2)は、中級下で「たら」の頻度が最も多い中国語母語話者 CIL03 のデータで、中国の西安と京都との違いについて質問している会話の一部である。

(2) T：うんうんうん、じゃセーアン、と京都はどう違いますか
S：んー、、セーアン、〈うん〉人口多いです、〈うん〉うん、あとで、んー、きれいな車少ないな、〈んーふーん〉{笑い}、〈んー、、うん〉そうですね、、うん
T：京都は
S：京都はー、〈ええ〉、んー、、車いっぱいなあ、〈うんうんうんうん〉本当に、これはびっくりしましたにほん来**たら**、〈うーん〉どこでも車いっぱいなー、自転車少ないなー、〈んー、うん、うん〉んー、〈う、なるほど、よくわかりました｛笑い｝〉んー、でも、お寺いっぱいですセーアンも同じです、〈んーんー、うんんー〉お寺いっぱいです
（KY コーパス，CIL03　注：「たら」の太字・下線は筆者による。以下同じ。）

(1)では、学習者の発話が単語レベルであったのに比べ、(2)では複文も使って発話がなされている。この 2 名の学習者はともに中級下のレベルで判定されているが、発話の量も質もかなり異なることが分かる。(1)の学習者は単語レベルの発話が多いため、語数が少なく、条件節を作る「たら」も出現しない。「たら」の出現には複文が発話できる能力が必要であり、その能力があれば逆に語数は一定以上の量になる。(2)のように、一定以上の語数があってはじめて出現する「たら」の調整頻度を算出するのに、(1)のような「たら」が出現しない語数の少ないデータを合計して計算するのは原理的に問題がある。「たら」の調整頻度は、「たら」が出現した学習者の語数で割ってこそ意味を持つ。

調整頻度は、各学習者の語数が同じなら個別平均と同じ値になる。この場合、個々の

学習者ごとに調整頻度を算出する手順を省いて、レベル別の合計で計算しても問題はない。しかし、個々の学習者によって語数が異なる学習者コーパスの場合、合計を用いて計算しても正しい値にはならない。個々の学習者ごとに個別調整頻度を算出し、その平均を求めるのが適切な方法である。

個別平均は個別調整頻度を使用して算出した平均値である。しかし、この平均値も妥当な代表値であるとは考えにくい。中級下では 9 名中 6 名の学習者の頻度が 0 だが、1 名の学習者が 5.1 回使用しているため平均値が 0.8 回になっている。平均値はデータの中心的傾向を代表した値だが、1 名の学習者のためにその傾向が歪んでしまっている。平均値で比較する際の統計的な適用条件は、データが正規分布に従い、比較するデータ同士の分散が一致していることである。データ量が少なく外れ値が多い学習者コーパスは、この条件に合いにくい。

中央値には調整頻度や平均値に見られるような適用条件の問題はない。しかし中央値は学習者の調整頻度を順番に並べた時の真ん中の学習者の頻度情報しか使用していないため、小規模のデータでは偏った値になる場合がある。表 5.2 の中央値割合を見ても中級上で 156%、上級で 70%と不安定な値を示しており、これが全体の中心的傾向を示しているかどうかは、全体の分布を確認しないと判断が難しい。

使用者数割合も原理的な問題はなく、情報を単純化しているだけに分かりやすい。しかし使用者数割合は、「たら」を 1 回でも使用した学習者は、「たら」を習得していると見なす考え方である。これは 1 回の使用も 10 回の使用も同一視するため、「たら」の習得を甘い基準で認定してしまう可能性がある。中央値や使用者数割合は、本来利用できる情報の一部しか使用しないため、取得した情報量を生かし切れていない。

第 1.4 項　代表値と分布図を併用した頻度比較の方法

調整頻度や平均値は、算出の原理や適用条件に問題がある。中央値や学習者数割合は、情報の一部しか有効活用できていない。これらの代表値が有効に機能しないのは、使用したデータの性質によるところも大きい。KY コーパスはデータ量が少なく、学習者ごとのばらつきが大きいデータである。そのようなデータを使用しているにも関わらず、データが正規分布に従っている時に最も威力を発揮する平均値を算出しても、代表値が有効に機能しないのは当然であろう。

調整頻度や平均値などは、データを一つの値に要約する代表値である。代表値はデータの特徴を端的に表すため多くの分析で使用されている。しかし、学習者コーパスのよ

うにばらつきの大きいデータの場合、学習者の分布を全体的に俯瞰する分析法の方が適していると思われる。これには、分布図による観察が有効である。第 4 章では、文書ごとの語数が異なるコーパスにおける分布観察の方法として、語数と単語頻度による散布図と、個別調整頻度を使用した度数折れ線の観察を行った。本章でもこれらの方法を使用して分析する方法が考えられるが、KY コーパスの場合、学習者数が少なすぎるため、度数折れ線による頻度比較が有効に行えない。そこで、本章では散布図に絞って分布観察の方法を検討する。

まず、はじめに学習者全員のデータを使用して学習者の語数と「たら」の頻度の散布図を描き、データに異常がないかどうかを観察するとともに、大まかなデータの特徴を把握することにする。図 5.1 は、粗頻度を使用した語数と単語頻度の散布図である。

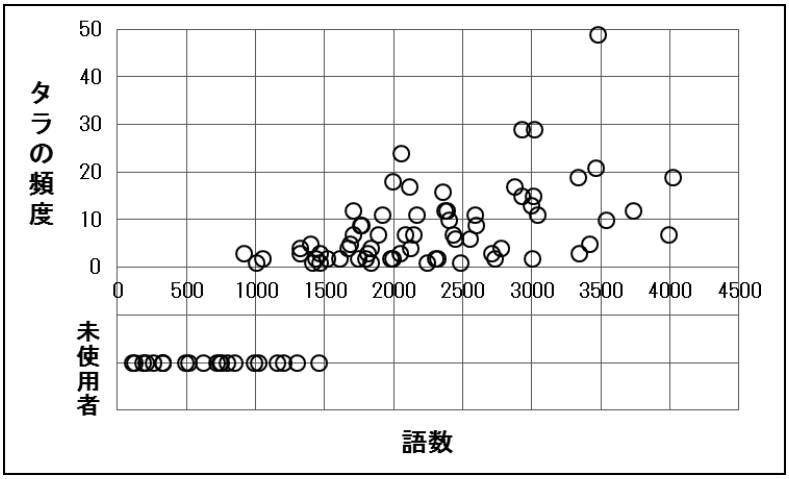


図 5.1　KY コーパスの学習者語数と「たら」の粗頻度（頻度 0 の未使用者は下段に表示）

図 5.1 では、未使用者は頻度 0 の位置ではなく、分布が分かりやすいように下段に図示した。語数と頻度の積率相関係数は.618 でやや強い相関がある。KY コーパスで語数が多いということは、30 分以内のインタビューで多くの発話を行ったことを意味しており、基本的に習得レベルの高い学習者ほど語数が多い傾向がある。このため、語数と「たら」の頻度にやや高い相関があるということは、学習者の習得レベルが高くなるほど「たら」が使用されている可能性を示唆している。しかし、図 5.1 を見るとばらつきも多く、語数が最も多い 4,000 語を発話した学習者の 2 人の「たら」の頻度が最も高くなっているわけではない。

「たら」を最も多く使用している学習者は超級の韓国語母語話者の KS09 で、3,500 語の発話で 49 回「たら」を使用している。次の(3)はその用例の一部である。

(3) T：じゃあ、あのー、Sさんのね、あのふつうの日、授業がある日、〈はい〉の、いちにちーをちょっとー、朝起きてから、え、をー簡単に教えてもらえませんか

S：えっと一応ー、まあ授業があるー30 分前におきて、〈うんうん〉顔洗って、〈うん〉歯磨いて、〈うん〉服着てから、出て、いえを出て、〈ん〉んで学校に着いてからまあ、5 分余った**たら**、〈うん〉余ってる時間ジュースをちょっと飲んで、〈うん〉で、授業に入って、〈うん〉でまあそれでー1 講目 2 講目ー終わった**たら**、昼ー食べて、〈うん〉で 3 講目入って 4 こ目、終わった**たら**まあ 4 こ目終わってー、まあ、1 時間小 1 時間ぐらいー、えーっとリブレとか、〈うん〉あのー学食にいって、〈うん〉ちょっと食べって、〈うん〉友達としゃべって、〈うん〉んでーそれからー、まあ帰ってくるんですけれど [·····]

T：うんー、あのーどんな番組を見てるんですか今

S：えーとー今はまあ、帰った**たら**、結構夜遅くーなるんで、〈うん〉洋画とか、〈うん〉じゃなかった**たら**、ああ、欠かさずに毎日見てんのは週間天気予報とか、 （KY コーパス、KS09）

(3)の用例中には「たら」が 5 回出現しており、頻度的には多いが、何らかの特殊要因のために多くなっているわけではない。この他の部分でも自然な会話が続いており、データに異常があるために「たら」が多くなっているという印象はない。

一方、「たら」の未使用者には明確な特徴があり、未使用者は語数が少ない学習者に限られている。1,500 語を超える語数を発話している学習者の中で、「たら」の未使用者は一人もいない。逆に 1,000 語以下しか発話できなかった学習者の中で「たら」が発話できた学習者はほとんどいない。この結果からも、「たら」の調整頻度を習得レベルの合計値で計算する従来の方法には、原理的に問題があることが裏付けられる。

次に、習得レベル別に描いた学習者の散布図を比較する。図 5.2 は横軸を学習者のレベル、縦軸を「たら」の個別調整頻度にし、統計ソフトの R を使用して作図した散布図である。母語には名大会話コーパスのデータを使用している。図 5.2 では、用例(3)のデータ（超級の頻度 15 付近のデータ）が、他の学習者と比較して非常に多くの「たら」を使用していることが分かる。スクリプトを読んで表現内容を観察する場合、そこに特に問題がなければ外れ値とは認めがたい。「たら」の頻度が高いという印象はあっ

ても、頻度が高いだけでは、異常な文章とはいえない。しかし、学習者の分布を描くと、他の学習者と比べてどれぐらい異なっているかの比較が視覚的に可能になる。図 5.2 を見ると、この学習者のスクリプトは、他のデータ集合とは大きく値が異なる外れ値となっていることが分かる。中級上にもこのような外れ値が見られる。個別調整頻度を算出することで語数が異なる学習者の比較を可能にした上で、習得レベル別の散布図を描くと、用例観察では判断しにくかった学習者発話の特殊性がよく分かる。

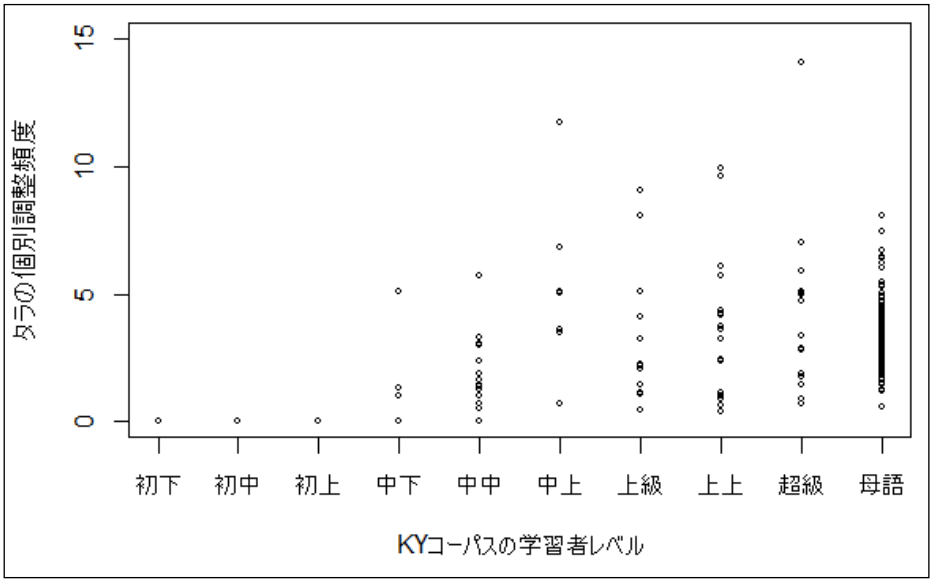


図 5.2　学習者レベルと個別調整頻度の散布図

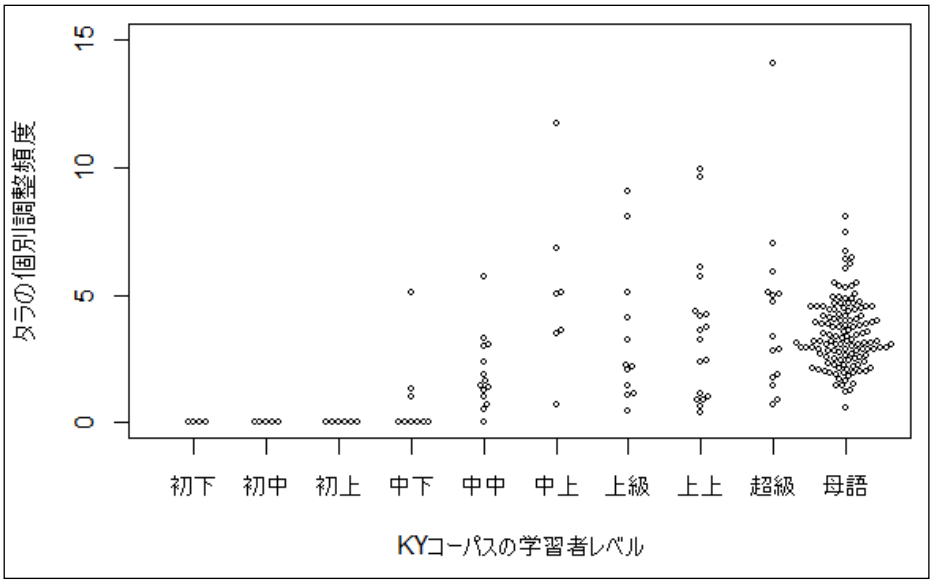


図 5.3　学習者レベルと個別調整頻度の蜂群図

ただし図 5.2 のような一般的な散布図では、一つのマーカーに何名の学習者が重なっているのかよく分からない。このため、R で beeswarm のパッケージを使用し、マーカーが重ならないように蜂群図で描いたのが図 5.3 である。母語のマーカーが多いのは、会話数が 129 の名大会話コーパスを使用したためである。図 5.2 では頻度 0 のマーカーが重なり、学習者が何名いるのか分かりにくいが、図 5.3 ではマーカーが横に広がっているため、把握しやすい。

しかし図 5.2、図 5.3 では、レベルが上がるにつれ、「たら」の頻度が直線的に上昇しているように感じられる。これは、無意識のうちにグラフの頂点付近のマーカーをつないで観察してしまうからである。グラフの頂点付近には外れ値も含まれるため、このような観察は好ましくない。外れ値の影響を受けないグラフを描くには、縦軸を頻度ではなく順位に直して作図する方法が考えられる。しかし順位に直すと、せっかくの頻度情報を失ってしまうため、図 5.4 のようにデータの分布比率を描画できる箱ひげ図を重ね書きする方法の方がより優れている。

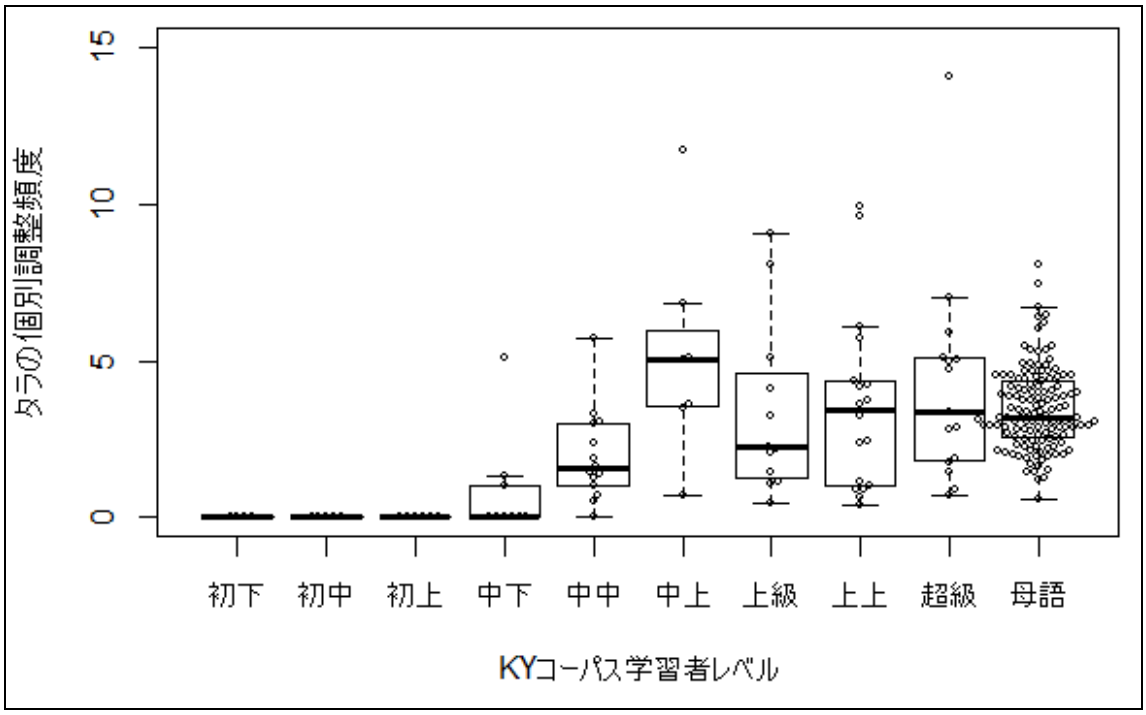


図 5.4　学習者レベルと「たら」の個別調整頻度の合成図。母語：名大会話コーパス

箱ひげ図は横の太線が中央値、箱の下限が第 1 四分位点、上限が第 3 四分位点を表す。つまり中央値を挟んで、25%から 75%の学習者が箱の中に納まる。ひげの長さは中央値から箱の長さの 1.5 倍までとすることが多く、このひげの外側のデータは外れ値と見な

すのが一般的である。このように箱ひげ図は、中央値という代表値を中心に、5 種類の要約値を使って大まかな分布傾向が表示できるグラフである。図 5.3 では、頂点の推移に目が向きやすいが、図 5.4 なら中央値という中心的傾向の推移や箱の大きさによるデータのばらつきの具合、外れ値の見極めなどが容易で、レベルごとの比較がしやすい。また、単なる箱ひげ図とは異なり、学習者の値がマーカーで示されるため、箱の位置や大きさが何に基づいて描かれているのか理解しやすい。

図 5.4 では、中級下から上級上にかけて、「たら」の使用が徐々に増加し、やがて頭打ちになっていく様子が観察できる。ただし、中級上だけは他のレベルと大きく傾向が異なっている。これはなぜであろうか。表 5.2 で、中央値割合を比較した際は、中央値割合が 50%→156%→70%と大きく揺れることしか分からなかったが、図 5.4 では、中級上で最も頻度の低い学習者は 1 回程度、他の多くの学習者は 5 回前後、最も多い学習者は 12 回程度と、学習者によって大きなばらつきがあることが見て取れる。このような違いがなぜ生じるのか、学習者の用例で確認してみよう。

次の (4)は、「たら」の個別調整頻度が最も多かった中国語母語話者の発話の一部である。ここでは日本の大学生が中国の大学生に比べて授業を欠席しがちであり、先生も注意しないことに戸惑いを覚えているという S に対し、T が質問している。

(4) T：あなたが大学の先生だっ**たら**どうしますか
S：だっ**たら**、〈うん〉そうですねー、日本へ、来る前、〈ええ〉ちゅ、たぶん厳しい先生、〈ええ〉自分が先生だっ**たら**、学生いなかっ**たら**、ちょっと気持ち悪い、〈んー〉でも日本へ来て、{笑}今、そこで、日本の教育、制度、ちょっとだけ、わかるようになっ、なったので、〈はい〉んー、帰っ**たら**、先生た、になっ**たら**、どうにかなる、まだわかりませんでした、〈はー〉わかっておりません、たぶん中国の先生より、厳しくない、〈うん〉日本の先生より厳しい、〈あー〉その間かもしれません
（KY コーパス，CIH01）

次の(5)は、「たら」の個別調整頻度が中央値に近い、5.1 の学習者の会話である。ここでは、バスケットボールのルールの説明を求められている。

(5) T：あー、バスケットボールは、見たことはあるんですけど、わたしはルー

ルがよくわからないんですけども、などういうふうになったら、勝つんですか

S：やー、バスケットボールは、それはやっぱり、あのー、点数が高くなって、〈ええ〉勝ちます、あの、自分の、あの、自分、いや、あれは、なんという、わからないです、あの、ボールは、自分の、方に、なげて、〈ええ〉あのー、あたったら、あの、点数が、とります

T：あたったら

S：はい

T：どこにあたるんですか

S：あの、バスケットボールは、あの丸いの、〈ええ、丸い〉日本語、何という分からないです

T：んーどんなものですか

S：たかいのものは、〈はい〉あのー、ボールをそのなかに投げて、〈ええ〉もし当たる、入れます、〈はい〉あの、ボールは、いれ、たら、点数が、あげます　（KY コーパス，CIH02）

(4)や(5)では、面接官の「たら」を含んだ質問に答えるために、学習者にも「たら」が多く使用されている。これに対し次の(6) は「たら」の粗頻度が 1 回と、中級上で最も少なかった英語母語話者の会話の一部である。ここでも学習者はトライアスロンの内容や筋肉トレーニングなどのスポーツの説明を求められているが、面接官に「たら」の使用がなく、話題も仮定の内容を含まないため、引用部分には「たら」の使用が見られない[16]。

(6)　T：トライアスロンって何ですか

S：トライアスロンは、バイクとか、〈はい〉水泳とか、〈はい〉running、とか、〈ええ〉ずっと、race、全部で、〈はい、ふーん〉race、ですね、長い、とても、大変、〈ふーん〉ですね

[16] タグ付き KY コーパスで検索するとこの後の会話で「筋肉だめ、〈あー〉だったら、」とタラが 1 回出現するが，山内博之氏から提供を受けたＫＹコーパス version1.2 のスクリプトでは「筋肉だめ、〈あー〉だから、」となっており，タラは使用されていない。両者で表現が異なる理由は不明である。本稿の分析はタグ付き KY コーパスの検索結果を使用しているため，EIH03 はタラを 1 回使用していると見なした。

T：それを、まあやってらっしゃるんですね

（中略）

T：うん、私、実は、そのウェイトトレーニングっていうのを、よく知らないんですが、〈うん〉あのー、具体的にはどんなことをするんですか

S：あのねー、いろいろ、あの筋肉は、いついつ全部違いますね、〈ええ〉だから special、練習あります、〈あーん〉この１つずつ、〈えーえー〉だから、あの、この、この、あの、筋肉は、こんな練習、だから、こんな筋肉は、こんな練習、〈うん〉全部違いますから、〈あーそうですか〉うん

T：でも、いつも、あのー、決まったことをやっているんですね

（KY コーパス，EIH03）

中級上 7 名のスクリプトを観察すると、面接官の「たら」の使用に合わせて学習者が「たら」を使用しているケースが多く見られる。学習者の「たら」の使用には一定の負荷がかかっている可能性が高いため、これらの頻度分布は例外と見なすのが妥当だと思われる。

中級上のデータを例外と見なすと、「たら」は中級下から一部の学習者で使用が始まり、中級中から徐々に使用が多くなり、上級上で頭打ちとなる。超級の分布は母語話者の分布とよく似ており、超級では母語話者と同じレベルで「たら」が使いこなせるようになっていると考えられる。表 5.2 のような代表値ではこの変化の詳細が分からないが、すべての学習者を一つのグラフ上にプロットできる蜂群図と、その統計的な分布の比率を図示できる箱ひげ図との合成図を使用すれば、学習者は大きなばらつきを持ちながらも、一定の傾向性を持って「たら」を習得し、母語話者と同レベルになっていく状況が的確に比較できる。この合成図を使用すれば、代表値と散布図のよさをともに生かした分析が可能である。

第 2 節　頻度分析法の比較：I-JAS の場合

前節では KY コーパスを使用して、「たら」の検索データを用いた代表値の比較を行った。しかし KY コーパスは総語数約 17 万語の小規模コーパスである。このため前節の結論は小規模コーパスでのみ成り立つ結論に限定される恐れがある。そこで本節では、I-JAS の第一次～第三次データを使用して同様の比較を行い、前節で得られた結論が大規模なコーパスでも変わらないことを示す。

第 2.1 項　使用するデータの説明

I-JAS は、国立国語研究所によって構築が進められている学習者コーパスで、言語類型論における分類を参考に決定された 12 言語の母語話者 1,000 名に対し、インタビューをはじめとした 8 種類のタスクを行ってデータが集積されている（迫田（編），2016；迫田・小西・佐々木・須賀ほか，2016)。また、各学習者の学習歴、留学経験、日本語環境、日本語を始めた動機などの詳しい個人情報が調査されている。

表 5.4　I-JAS の調査実施国・機関と人数および公開期

迫田・小西・佐々木・須賀（2016）p.18.表 2、p19.表 3、p.20.
表 4、p.21.表 5、p.200.表 10 をもとに筆者が合成して作成した。

	実施国	調査実施機関	母語	調査ID	人数	公開期
1	インドネシア	インドネシア教育大学	インドネシア語	IID	50	1・4
2	スペイン	マドリッド・コンプルセンテ大学	スペイン語	SES	50	1・3
3	タイ	タマサート大学	タイ語	TTH	50	1・4
4	トルコ	ボアジチ大学	トルコ語	TTR	50	1・2
5	オーストリア	ウィーン大学	ドイツ語	GAT	50	1・4
6	ドイツ	ハイデルベルグ大学	ドイツ語	GDE		
7	ハンガリー	カーロリ・ガーシュパール・カルビン大学	ハンガリー語	HHG	50	1・3
8	フランス	グルノーブル スタンダール大学	フランス語	FFR	50	1・4
9	ベトナム	ホーチミン市師範大学	ベトナム語	VVN	50	1・3
10	ロシア	モスクワ市立教育大学	ロシア語	RRS	50	1・3
11	アメリカ	アリゾナ州立大学	英語	EUS	100	1・2・5
12	イギリス	オックスフォード・ブルックス大学	英語	EGB		
13	オーストラリア	ニューサウスウェールズ大学	英語	EAU		
14	ニュージーランド	オークランド大学	英語	ENZ		
15	韓国	高麗大学	韓国語	KKR	50	5
16	韓国	東国大学	韓国語	KKD	50	1・2
17	中国	上海海事大学	中国語	CCM	50	1・2
18	中国	湖南大学	中国語	CCH	50	4
19	台湾	東呉大学	中国語	CCS	50	5
20	台湾	台中科技大学	中国語	CCT	50	3
21	日本	広島YMCA・イーストウエスト日本語学校・東京日本語学校	混在・教室環境	JJC	100	1〜5
22	日本	広島YMCA・広島国際学院大学・広島文化学園大学	混在・教室環境	JJE		
23	日本	広島YMCA・公益財団法人東広島市教育文化振興事業団・新宿あけぼの会	混在・自然環境	JJN	50	1〜3
24	日本	国立国語研究所・A社会福祉法人施設	日本語母語話者	JJJ	50	1・2

I-JAS は、2016 年 5 月の一次公開に続き、2017 年 5 月の第二次データの追加によって学習者 400 名、日本語母語話者 50 名に、2018 年 5 月には第三次データの追加によって学習者データが 610 名になった。。最終的に 5 期に分けて公開が行われる予定で、全てが公開されるのは、2020 年 3 月頃が見込まれている。表 5.4 は迫田・小西・佐々木・須

賀ほか（2016）の複数の表から、関連する情報をまとめて集約した表である。

表 5.5 は、第一次～第三次データの I-JAS 語数表[17]を使用して学習者 610 名と母語話者 50 名について、タスク別の平均語数をまとめた表である。

表 5.5　I-JAS のタスクと第一次～第三次データの学習者・母語話者の平均語数

記号	タスク	学習者平均	SD	母語話者平均	SD	学習者語数÷母語話者語数
I	インタビュー	2,264.2	740.6	4040.9	1059.1	56.0%
RP1	ロールプレイ1	194.6	77.2	216.3	81.3	90.0%
RP2	ロールプレイ2	193.6	84.8	219.0	85.3	88.4%
ST1	ストーリーテリング1	123.6	58.1	116.7	44.7	105.9%
ST2	ストーリーテリング2	138.4	62.5	125.3	38.1	110.4%
D	絵描写（232名のみ）	324.9	135.0	369.3	158.9	88.0%
SW1	ストーリーライティング1	92.1	32.9	98.9	25.1	93.2%
SW2	ストーリーライティング2	96.9	34.4	112.1	28.4	86.5%
	学習者別タスク合計	3,122.0	976.5	5298.5	1189.5	58.9%

データ量はインタビューが最も多く、全体の約 7 割を占めている。インタビューの時間は約 30 分で、日本語学習の動機や好きな本など現在の状況を聞く内容、幼少期の体験や恩師の話など過去の体験を聞く内容、将来の夢など未来のことを聞く内容、および「お金と時間とどちらが大切か」などの意見を聞く内容などが含まれている。

ロールプレイは週 3 日のアルバイトを 2 日に短縮してもらう依頼と、ホール係から調理係に持ち場を変更するように依頼されて断る内容である。ストーリーテリングは 4 コマと 5 コマの漫画を見ながらその内容を語るタスク、絵描写は許（1997）で使用されたアスペクト表現を産出させるための絵を見て、絵の内容を発話するタスクである。このタスクだけは、調査の実施途中から採用されたため、第一次～第三次データの場合では 610 人中 409 名だけが実施している。ストーリーライティングは、ストーリーテリングと同じ漫画を見て、作文を書くタスクである。

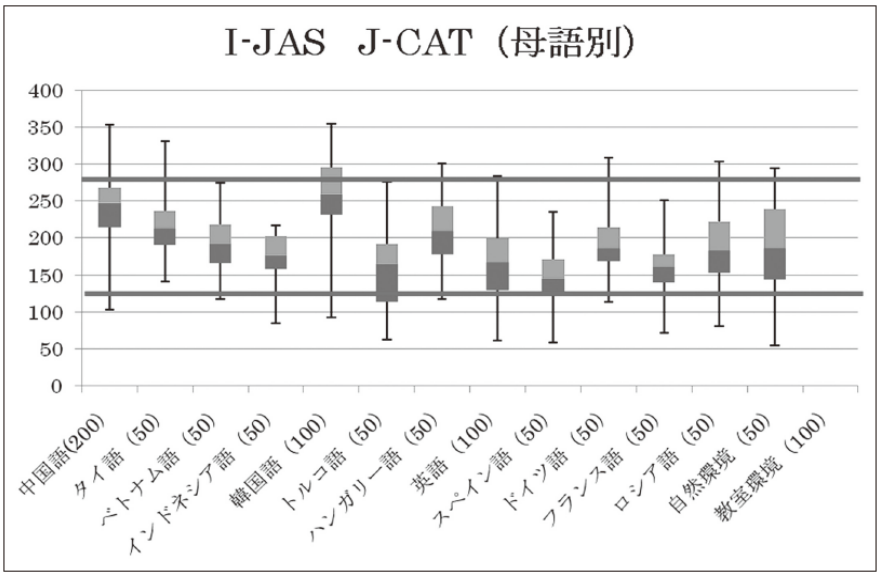
これらのタスクにおける学習者と母語話者との語数の違いはインタビューが最も大きく、それ以外のタスクではそれほど大きな差はない。この理由は、インタビューに比べ、他のタスクに必要とされる語数が少なく、一定の課題をこなすためには、母語話者と同程度の言語量が必要であったからだと思われる。

このように I-JAS では複数のタスクによって複数の文書が集積されているが、第 2 章

[17] 中納言バージョン 2.4 2.4.2　短単位データ 20180502 版　I-JAS 語数表_20180502 版.xlsx

で述べたとおり、これらの文書は学習者の言語的特徴をよりよく反映させるために、多方面からデータを集めた結果であり、I-JAS の観察単位は学習者と考えるのが妥当である。

I-JAS では各学習者に J‐CAT（Japanese Computerized Adaptive Test）の聴解、語彙、文法、読解の各 100 点満点中の得点と合計点、および SPOT（Simple Performance‐Oriented Test）の 90 点満点中の得点が付与されている。図 5.5 は、迫田・小西・佐々木・須賀ほか（2016:199）より引用した母語別の J-CAT 得点の箱ひげ図である。これを見ると、母語によって J-CAT 得点が大きく異なっていることが分かる。



迫田・小西・佐々木・須賀ほか，2016:199 図 20 を引用

図 5.5　母語別の J-CAT 得点の箱ひげ図

図 5.5 の分布が、母語別の一般的な学習者レベルの分布を表しているのか、あるいはたまたま選ばれた調査機関の学習者レベルに強く依存しているのかは不明である。図 5.5 を見ると、中国語や韓国語の母語話者の成績が飛びぬけてよい。これは漢字圏の学習者や日本語の文法に類似している母語の学習者であることを考えると、一般的な傾向であるようにも思われる。しかし、その次に成績のよいグループがタイ語とハンガリー語である。ハンガリー語の場合、日本語と同じ膠着語であるが、タイ語には日本語との共通点は少なく、タマサート大学という調査機関で学ぶ学習者が優秀であったために J-CAT 得点が高くなっている可能性が考えられる。I-JAS では、母語別に学習者が無作為抽出されたわけではなく、限られた機関から便宜的に学習者が選抜されているため、

その分析結果が何を表しているのかについて判断することが難しい。このような問題は学習者コーパスの構築に当たってはある程度避けられない問題だが、I-JAS の母語別の比較については慎重に解釈を行う必要がある。コーパスを使用した分析を行う場合には、異なった設計方針によって製作された複数のコーパスを使用し、これらの結果を比較することで結果の一般化を行うことが重要である（スタッブズ，2006:313）。学習者が無作為抽出されていない学習者コーパスにおいては、特にこの点に留意する必要がある。

本節では OPI をもとに製作された KY コーパスと比較するため、I-JAS の第一次～第三次データにおける発話データ約 216 万語を使用して分析を行う。I-JAS も名大会話コーパスと同様に、国立国語研究所が開発した検索インタフェース「中納言」による検索が可能になっている。本節ではこれを利用し、語彙素「た」、品詞「助動詞」、活用形「仮定形」で検索した「たら」4,368 語のデータを使用する[18]。

I-JAS では各学習者に J-CAT の聴解、語彙、文法、読解の各 100 点満点中の得点と合計点、および SPOT の 90 点満点中の得点が付与されている。したがって、OPI のレベルと J-CAT や SPOT の得点の対応付けができれば、基本的にはレベルを合わせた比較が可能になる。ただしこれらは異なった枠組で作成されたテストであるため、機械的な対応付けができるわけではない（李・小林・今井・酒井ほか，2015:53-4）。本章では、J-CAT のスコア互換表[19]を利用し、J-CAT の合計点を KY コーパスのレベル区分に読み替えて分析を行う。区分は 100 点以下：初級、101～150：中級下、151～200：中級中、201～250：中級上、251～300：上級、301～350：上級上、351 以上：超級の 7 種類とする。この区分はあくまでも大まかな目安であり、この区分と OPI のレベルにどれぐらい密接な対応があるのかは、今後多くの学習項目に基づく比較を通して検証していく必要がある。

第 2.2 項　代表値を使用した頻度比較の結果

表 5.6 は I-JAS 第一次～第三次データを使用して、KY コーパスの表 5.2 と同様の方法で集計した分析結果である。学習者数の欄を見ると、I-JAS では初級と上級上、超級の人数が少なく、中級中と中級上に学習者が集中しているのが分かる。

表 5.6 で①調整頻度と②個別平均を比較すると、KY コーパスと同様に少しずつ値が異なっている。学習者数が 187 名の中級中でも、両者の値は一致しないことが確認され

[18] 2018.10.05 閲覧。

[19] http://www.j-cat.org/page/interpret、2017.01.07 閲覧。

る。②個別平均と③中央値は、上級以下のレベルで値が大きく異なっている。この理由は、平均値を算出する際、外れ値を含んだまま計算したため、外れ値に影響されて平均値がデータの中心的傾向を示さなくなったことによる。表 5.6 を見ると、学習者数が多くなったからといって、各代表値が必ずしも安定するわけではないことが確認できる。

表 5.6　I-JAS の基礎統計量と「たら」頻度に基づく代表値の分析結果の比較

				⓪	①	②	③	④	①’	②’	③’	④’
レベル	人数	語数合計	語数平均	タラ粗頻度	調整頻度	個別平均	中央値	使用者数	調整頻度割合	個別平均割合	中央値割合	使用者割合
初級	30	63,176	2,106	60	0.9	0.8	0.0	13	39.7%	32.2%	0.0%	43.3%
中級下	87	244,853	2,814	331	1.4	1.2	0.3	45	56.6%	49.2%	12.5%	51.7%
中級中	187	549,609	2,939	843	1.5	1.4	1.0	150	64.2%	58.8%	45.1%	80.2%
中級上	195	645,013	3,308	1374	2.1	2.1	1.5	181	89.1%	86.5%	67.4%	92.8%
上級	90	326,202	3,624	910	2.8	2.6	2.2	86	116.7%	109.3%	100.1%	95.6%
上級上	19	73,484	3,868	205	2.8	2.7	2.7	18	116.7%	113.3%	119.0%	94.7%
超級	2	7,652	3,826	37	4.8	4.7	4.7	2	202.3%	195.8%	210.4%	100.0%
母語	50	254,381	5,088	608	2.4	2.4	2.2	49	100.0%	100.0%	100.0%	98.0%

第 2.3 項　調整頻度の妥当性

本節では I-JAS 中級中のデータを使用し、調整頻度と個別平均の値が食い違う理由を考察する。この理由は図 5.6 を見ると理解しやすい。図 5.6 は中級中の語数と「たら」の個別調整頻度の散布図である。

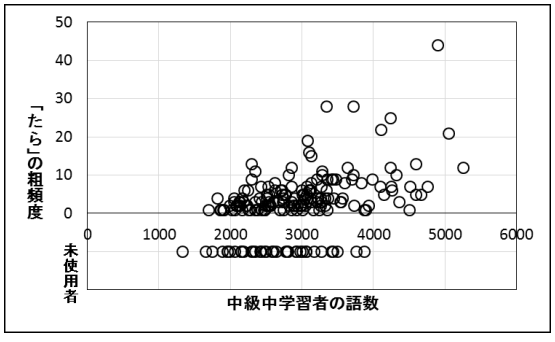


図 5.6 中級中の語数と「たら」頻度の散布図

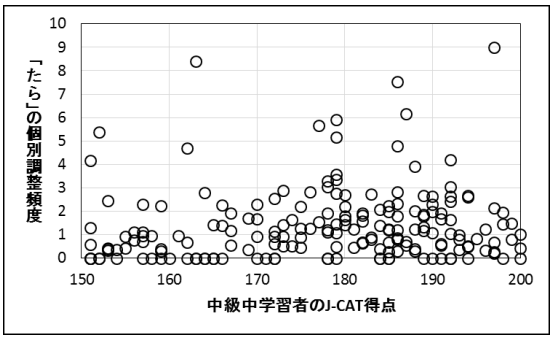


図 5.7 J-CAT 得点と「たら」頻度の散布図

この図で、「たら」の頻度が 0 の未使用者（図 5.6 下段）に着目すると、「たら」の未使用者は 4,000 語以下の学習者に限られることが分かる。この現象は、第 1.3 項で観察した KY コーパスと同一であり、「たら」の頻度を平準化するのに、語数が少ない傾向にある「たら」未使用者の語数を使用しているため、食い違いが生じると考えられる。

図 5.7 は J-CAT 得点と「たら」頻度の散布図である。これを見ると、J-CAT 得点と「たら」の頻度にはほとんど相関がない。得点が低くても「たら」を多用する学習者もいれば、得点が高くても「たら」を使用しない学習者もいる。中級は初級や上級に比べ多様な学習者が存在し、そのばらつきも大きいことが確認できる。このように発話の語数も「たら」の頻度も大きくばらついているデータを使用して、合計数で調整頻度を求めても、正確な値にはならない。頻度を調整するには学習者ごとに個別調整頻度を算出するのが適切な方法だと考えられる。

第 2.4 項　代表値の妥当性と合成図の有効性

本項では、合成図の観察を通し、平均値、中央値、使用者割合の妥当性と合成図の有効性を検討する。図 5.8 は、図 5.4 と同じ方法で描いた合成図である。I-JAS には日本語母語話者のデータが含まれているため、「母語」は学習者と同条件で採取した母語話者のデータを使用している。これに参考値として、図 5.4 で使用した名大会話コーパスのデータを加えた。

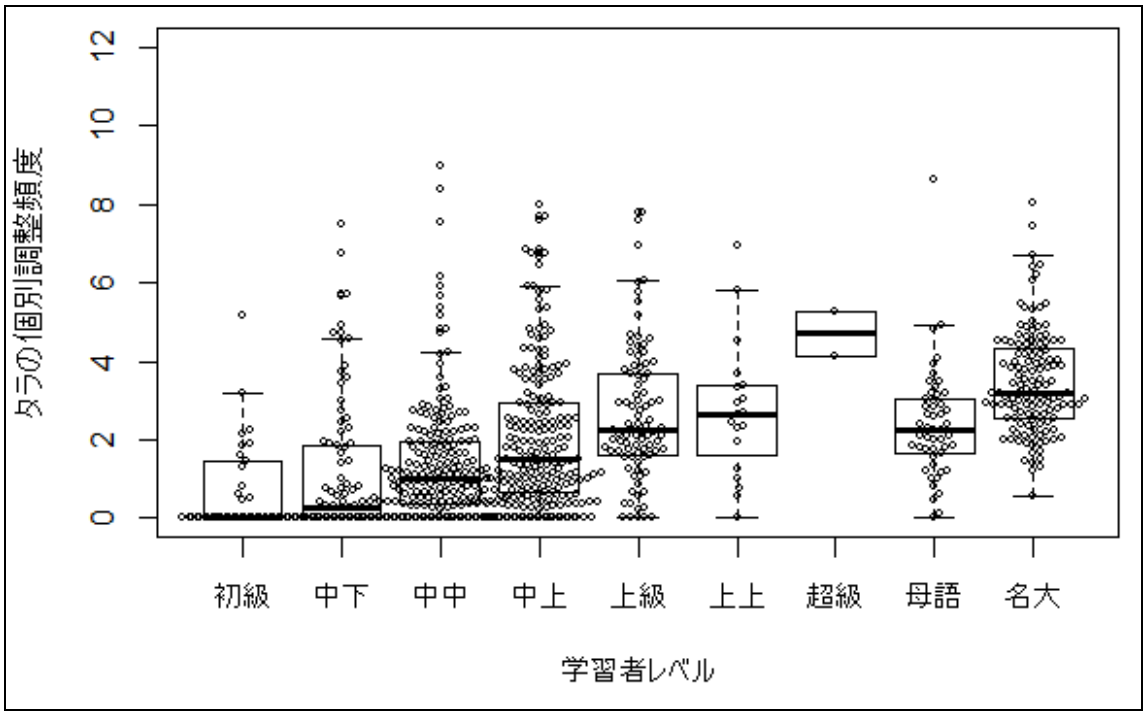


図 5.8 学習者レベルと「たら」の個別調整頻度の合成図

母語：I-JAS、参考：名大会話コーパス

図 5.8 は、それぞれの箱の大きさがコンパクトで、中央値も箱の真ん中に位置するものが多い。この理由は、I-JAS の場合、KY コーパスに比べて学習者数が多く、学習者

の語数も多いためだと思われる。中央値や使用者割合はもともと外れ値に強い代表値であるが、人数が増えたことによって、より代表値の信頼性が高まったと考えられる。ただし、当然ながらこの二つの代表値が限定した情報しか使用していないことには変わりはない。

一方の平均値は、データが増加したことによって信頼性が増したとは考えにくい。その理由は図 5.8 を見れば明らかなように、特に上級以下のレベルで多くの外れ値が存在するからである。平均値がデータの代表値としてふさわしいのは、母語のデータのように中央値を境に上下がほぼ対称になる場合である。母語でも外れ値が一つだけ見られるが、他のデータに強い正規性があるためほとんど影響はない。これに対して、学習者のデータはひげの長さも非対称で、ひげの上に多くの外れ値が続いている。この場合、平均値を使用しても中心的な傾向を表すことができない。中級に外れ値が多いのは、先に述べたようにこのレベルに多様な学習者が存在するためだと考えられる。したがって、いくらデータを増加させても、中級におけるばらつきは解消しない可能性が高い。このため大規模な学習者コーパスを使用しても、平均値や平均値の一種である調整頻度を使用することは妥当ではないと考えられる。

次に合成図の有効性を考察する。合成図を観察する利点の一つは、図 5.8 を観察することによって中央値の正確さや平均値の不正確さの理由が分かったように、グラフを観察することでデータの正確性が評価できる点にある。たとえば超級では、「たら」の頻度が飛びぬけて高いのは、超級にわずか 2 名の学習者しかおらず、たまたま頻度の高い学習者がそろったためだと思われる。その根拠を明確に示すのは難しいが、初級から上級上までの頻度増加の傾向から考えると、ここで超級だけ極端に頻度が増加するとは考えにくい。表 5.6 の代表値を使用した場合、算出された値が正しいという前提に立って分析するしかないが、合成図ではデータがどれほど正確であるかの評価も行いながら全体の分布を考察することができる。

このようなデータの正確性の評価を行いながら分析できる利点は、異なったコーパスの比較を行う際にも有効に働く。一般的にいって、コーパスはどのようなコーパスであれ、それぞれの設計方針に基づく何らかの偏りを持っている（スタッブズ, 2006:313）。このため単独のコーパスで、「日本語学習者」といった巨大な母集団の実態を推定することは困難である。しかも、それぞれのコーパスがどのような偏りを持っているかは、異なるコーパスを比較してみないかぎり分からない場合が多い。図 5.9、5.10 は図 5.4 の KY コーパスと図 5.8 の I-JAS の「たら」頻度を比較したグラフである。

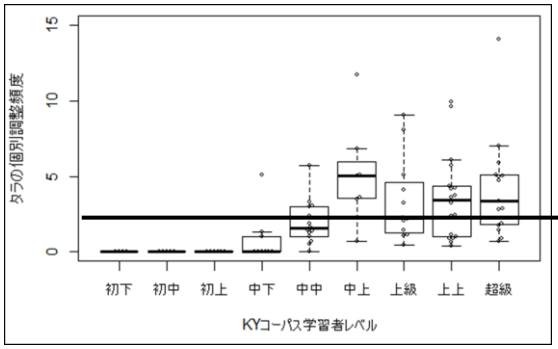


図 5.9　KY コーパスの「たら」頻度

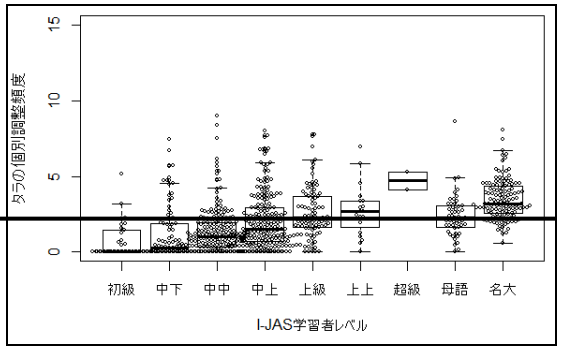


図 5.10　I-JAS の「たら」頻度

二つのコーパスを比較してまず分かることは、大きな傾向性が非常によく似ているということである。第 2.1 項では、OPI と J-CAT という異なる基準で区分されたデータを比較する難しさを述べた。OPI では単語や句のレベルでしか発話できないレベルを初級と定めているため、複文で使用される「たら」は中級以降にしか使われない。一方、J-CAT では聴解、語彙、文法、読解の総合点でレベル分けしているため、初級から「たら」を使用する学習者が観察される。このような違いはあるが、中級中から大半の学習者が「たら」を使用するようになり、上級以上では使用数がほぼ母語話者と同レベルになるという傾向は同じである。「たら」の使用頻度に限れば、OPI と J-CAT の区分を同列に使用しても、それほど大きな問題はなさそうに思われる。

問題は KY コーパスの中級上や I-JAS の超級のデータである。KY コーパスの中級上については、第 1.4 項でスクリプトを観察しながら、これがインタビューによって引き起こされた偏りである可能性を指摘した。しかしそれを本当に偏りといってよいかは判断が難しかった。ところが図 5.9 と図 5.10 で比較すれば、I-JAS の 195 名のデータに基づいて KY コーパスの中級上のデータは偏っていることが裏付けられる。I-JAS の超級データも、これが偏っているという判断は明確な根拠を示すのが難しかった。それが KY コーパスと比較することによって、15 名のデータに基づいて偏りを持っていることが裏付けられる。このような判断が可能になるのは、合成図に全学習者の情報がプロットされているため、データの前後のつながりや集積されている人数、そのばらつきの程度などを評価して判定できるからである。

さらに合成図では、この図にプロットされている学習者の情報を調べることで、分析をさらに進展させることが可能である。たとえば I-JAS では、20 か所の海外地域で調査が行われているだけでなく、国内の教室環境学習者と国内の自然環境学習者の調査も行われている。このため、学習者を JFL（Japanese as Foreign Language）と JSL（Japanese as

Second Language）に区分して分析することが可能である。

図 5.11 は図 5.8 の中から「たら」の個別調整頻度が 5 以上の学習者 39 名を抽出し、学習者の調査 ID を併記したグラフである。図で濃い網掛けをしているのが国内の自然環境学習者（JJN）16 名、薄い網掛けが国内の教室環境学習者（JJC と JJE）4 名である。第一次～第三次データ 610 名の中で JSL 学習者は、その 16.4%に当たる 100 名であるが、図 5.11 では、「たら」の個別調整頻度が 5 以上の学習者 39 名の半数に当たる 20 名が JSL 学習者で占められている。明らかに「たら」を多用している学習者には、JSL 学習者が多い。特に国内の自然環境学習者（JJN）は 50 名中 16 名が頻度 5 以上使用しており、習得レベルも初級から超級まで幅広い。国内の自然環境学習者には、会話で使用されやすい「たら」が習得されやすく、習得レベルが上がっても「と・ば」などの条件表現の習得が進まずに、「たら」が多用され続けている可能性が考えられる。

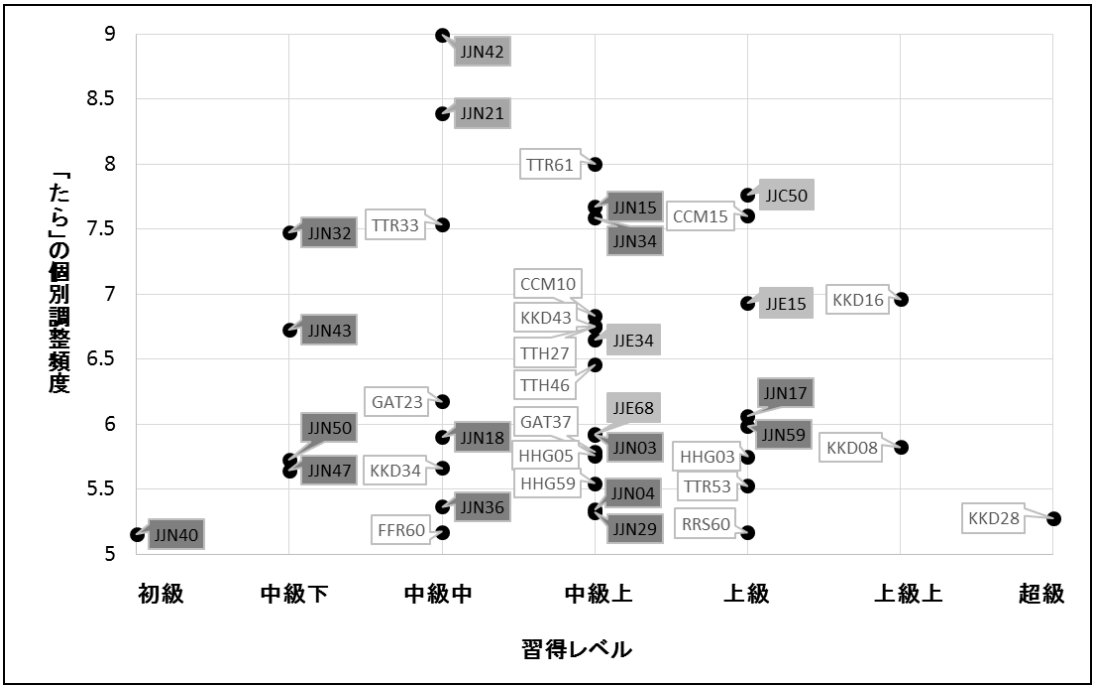


図 5.11　「たら」頻度 5 以上の学習者とその ID

JFL の中で特徴的なのは、韓国語母語話者が 5 名含まれていることである。このうち超級の KKD28 と上級上の KKD8 は、親しい友人に日本人がおり、仕事やアルバイトでも日本語を使用していることから、JSL に近い環境にいると考えられる。この 2 人を含め、上級上の KKD16 と中級中の KKD34 は、日本語でどのような活動しているかという質問に対して、テレビ・ドラマ・映画・アニメを見ると答えており、日本語の音声を

聞く機会が多い学習者が「たら」を多用している可能性がある。

平均値、中央値、使用者割合を使用した場合、「たら」の頻度がレベルによってどのように異なるかの比較はできるが、そこから分析を発展させていくことが難しい。合成図の場合、マーカーの一つ一つが個々の学習者に対応しているため、図 5.11 のような追加分析が可能になる。特に I-JAS では学習者ごとにフェイスシートが用意されているため、これを利用して、「たら」を多用する JSL と JFL に共通する要因をさらに追及していくことも可能である。

第 2.5 項　学習者コーパスにおけるデータ数と分布のばらつきの関係

前節前半では、学習者のデータが増加しても、平均値が妥当な代表値としては使用しにくいことを確認した。本節では母語話者データを観察することにより、この理由をさらに検討する。

I-JAS の母語話者データは、学習者と同じインタビューやタスクを行って収集されている。もう一方の母語話者データである名大会話コーパスは、雑談を収集したデータであり、集約した単位も発話者ではなく会話である。これらは全く異なったデータであるが、全体的な分布の傾向はよく似ている。図 5.12 はこれらを相対度数折れ線で描いた図で、両者ともほぼ正規分布に近似していることが確認できる。両者で頻度が異なる理由は、データが質的に異なることのほか、名大会話コーパスの調査地が名古屋を中心としていたことによる方言の影響が考えられる[20]。しかし、ここで注目したいのは両者が非常に狭い範囲で近似している点である。

図 5.13 は、I-JAS の 50 名の母語話者データから、ランダムに再サンプリングした 10 名、20 名、30 名の度数折れ線を描いた図である。図 5.13 はたまたま選んだデータを描いたに過ぎないため、一般化はできないものの、均質な母集団からランダムサンプリングしたデータは、基本的に元の母集団に類似した分布を示すことが観察できる。このうち 10 名のグラフの形が他の二つと異なるのは、データ数が少ないためだと思われるが、しかしそのグラフですら、正規分布に近い形をしている。図 5.12、図 5.13 から分かることは、日本語母語話者という均質な母集団から採取したデータは、10 名や 20 名といった少ない人数でも、正規分布に近似する可能性があるということである。その一方で、図 5.8 で観察したように、学習者のデータはばらつきが大きく、正規分布には従わない。

[20] 名大会話コーパスで使用されている言葉は共通語が大半を占めるが，問題は共通語を使用しているつもりでいながら，タラを多用している「気づかない方言」の可能性である（庵，2017:25）。

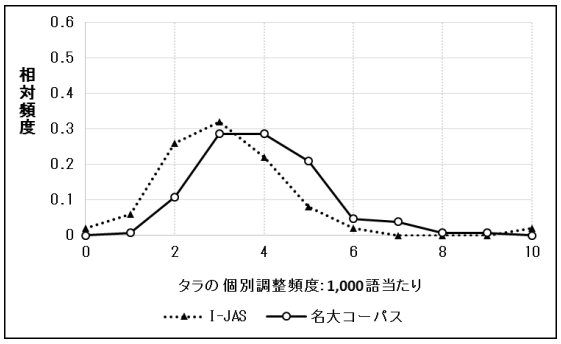


図 5.12「たら」頻度の度数折れ線・全データ

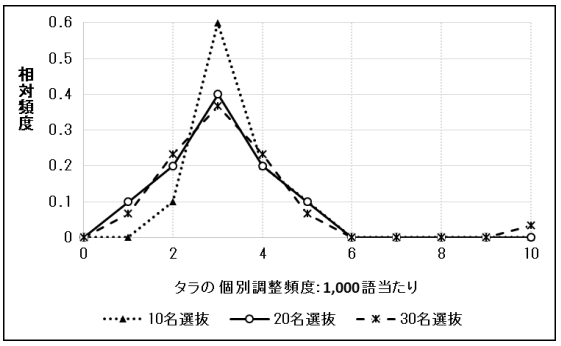


図 5.13 「たら」頻度の度数折れ線・I-JAS 選抜

その理由は学習者の「たら」使用が均質ではなく、さまざまな多様性を持っているためだと考えられる。つまり、学習者コーパスでばらつきが大きいのは、データ数の問題ではなく、学習者の多様性の問題である。このようなデータを分析する場合は、一つの値でデータを代表させるのではなく、ばらつきの状態を含めた全体の分布を観察し、どのような理由でデータがばらついているのかをさらに追求できる形で比較する方法が、有効な分析法だと考えられる。

第 3 節　まとめ

本章では、広義コーパスに属する学習者コーパスを対象に、代表値を使用した分析法を検討し、代表値と分布図を併用する分析法の提案を行った。均衡コーパスの分析で使用されている①調整頻度は、学習者ごとの語数が異なる学習者コーパスで使用すると不正確な値になる。レベル別の語数合計で頻度を平準化するのではなく、個別の学習者ごとに調整頻度を算出して分析するべきである。ただし、個別調整頻度を使用しても、②平均値は外れ値が多く、データの中心的傾向を示さない場合が多い。③中央値や「たら」の④'使用者数割合は原理的には妥当だが、一部の情報しか有効活用できないため、きめ細かな判断が難しい。

学習者のデータは個人によるばらつきが大きく、レベル別の習得状況を一つの代表値で点的にとらえるのは困難である。学習者の実態を面的にとらえるには作図による分析が適している。散布図であれば、学習者の全体の姿を一枚のグラフで表現できる。ただし、通常の散布図はマーカーの重なりが分かりにくいため、本章ではマーカーが重複しない蜂群図を使用し、さらにデータの四分位という代表値が視覚的に把握できる箱ひげ図を重ね書きする手法を採用した。蜂群図と箱ひげ図の合成図で観察すると、データの正確性を評価しながら全体の傾向性を分析できるだけでなく、個別のデータに着目した追

加分析も可能になる。図 5.11 で観察されたように、国内の自然環境学習者に「たら」の使用が多いという現象は、本研究の枠組みを使用することによってはじめて解明できた使用実態である。

数値を使用した分析を行う前に、データをグラフ化して観察するべきだという主張は、統計分析の基本であり、本章の分析もこれを支持する結果となった。本章の主張は、代表値を使用した分析法を否定することではなく、全体の分布観察に基づいて、使用するデータにふさわしい代表値を選択するべきだという点にある。箱ひげ図は、外れ値に強い中央値を中心に、5 種類の代表値を使用したグラフであり、散布図と併用することによってその値の解釈が容易になる。

散布図と箱ひげ図の合成図自体は、特に目新しい分布図ではないが、これまで学習者コーパスの分析にはほとんど使用されてこなかった。その理由は、これまで学習者コーパスの観察単位は文字、単語、文などの言語単位であると考えられてきたためである。学習者が観察単位であるという考え方に立たないかぎり、学習者の分布図を描くという発想は生まれない。その意味で、本章で提案した分析法の意義は、作図の方法より、学習者を観察単位として分析するという考え方自体にある。

第 6 章　カイ二乗検定の方法

本章では、コーパス言語学で最も多用されてきたカイ二乗検定を有効に行うための方法を考察する。コーパスのデータを使用したカイ二乗検定では、これまで主に二つの問題が指摘されてきた。一つはコーパスのような大規模データを使用すると、実質的には意味のない差でも有意になるという問題（Kilgarriff，1997；Oakes，1998:28-29）、もう一つは言語データには統計的検定の前提となる独立性が欠如しているという問題である（Kilgarriff，2005；Evert，2006）。本章では、これまでコーパス言語学で行われてきたカイ二乗検定の問題を検討し、統計学的にも言語学的にも有効なカイ二乗検定の方法を提案する。データには、BCCWJ 出版 SC 書籍レジスター（出版書籍）、特定目的 SC 白書（白書）、図書館 SC（図書館書籍）の短単位データを使用する。

第 1 節では統計的有意差に関する誤解と大規模データを使用した場合、効果量が有効に評価できない場合が多いという問題を指摘する。第 2 節では従来の方法を使用してカイ二乗検定のケーススタディを行い、第 1 節で指摘した問題点を具体的に観察する。第 3 節では言語データを使用したカイ二乗検定における独立性の問題を考察する。第 4 節では、コーパスにおける独立した個体は文書であるとする本研究の考え方に従ってカイ二乗検定を行い、文書を観察単位とした分析法の有効性を検討する。最後に第 5 節でまとめを述べる。

第 1 節　統計的検定における有意差と効果量の問題点

本節では、統計的に有意であれば言語学的にも有意義な差があると考えるのは誤解であること（第 1.1 項）、コーパスデータのような大規模データを使用した場合、効果量が有効に評価できない場合が多いこと（第 1.2 項）を指摘する。

第 1.1 項　統計的検定における有意差の誤解

統計的有意差と言語学的に有意義な差とは、基本的に無関係である。したがって、異なったコーパス間で、ある単語の使用率についてカイ二乗検定を行った際、検定結果が有意であれば使用率にも言語学的に有意義な差が存在すると考えるのは誤解である。統計的検定は標本で観察された現象が母集団でも成立するかどうかを確かめているに過ぎない。このため、ある単語の使用率の差が統計的に有意であるからといって、言語学的に意味のある差が存在しているとは限らない。特に統計的検定には「データ数が多ければ実質的にはほとんど意味がないと判断されるようなわずかな差でも統計的には有

意であることになってしまう」という性質が存在する（吉田，2001:240）。コーパス同士の語数が同じ場合、1,000 語のコーパスなら 4.4%の差、10,000 語のコーパスなら 1.4%の差で有意となる。数百万語を超えるコーパスであれば、実質的な差がほとんどなくても p 値は 5%水準で有意となる。このため BCCWJ のような大規模コーパスのデータを使用した場合、統計的検定を行う意味はそれほど高くない（吉田，2001:240）。

サンプルサイズが大きくなればなるほど実質的な差がない現象まで簡単に有意になりやすいという Kilgarriff（1997）等で指摘されている問題は、このような統計的検定の性質に根ざしており、基本的には有意義な差があるかどうかの評価を p 値ではなく、次に述べる効果量を使用して行うことで解決できる。

第 1.2 項　効果量とその評価基準の問題

効果量（effect size）とは、サンプルサイズによって変化することのない、標準化された効果の大きさを示す指標である（水本・竹内， 2008:58-9）。水本・竹内（2008，2011）では、欧米の学術誌において効果量の報告が義務化されている例などをあげながら、各種の統計的検定に対応した効果量の求め方とその効果量の判断の目安が啓蒙的に紹介されている。カイ二乗検定の効果量の一つであるクラメールの連関係数 V（Cramer's V ）の場合、American Psychological Association (APA) Publication Manual 第 5 版（2001）の記述に従って、効果量の目安は、小が.10、中が.30、大が.50 という水準が示されている（水本・竹内，2008:62，2011:51）。これ以後この目安を APA の目安と呼ぶ。これに従えば、コーパス間における使用率の差も Cramer's V が.10 以下の場合、実質的な差は小さいと考えられる。

（6.1 式）Cramer's $V = \sqrt{\dfrac{\text{カイ二乗値}}{\text{全データ数} \times (\text{分割表の行と列でカテゴリ数が小さい方の数} - 1)}}$

ところが日本語で書かれた代表的な言語統計分析の入門書である石川・前田・山崎（編）（2010:60-7）などで紹介されているカイ二乗検定の方法で Cramer's V を求めると、大規模コーパスのデータを使用した多くのケースで Cramer's V は.10 を下回るごく小さな値となり、効果量を有効に評価するのが難しくなる（石川，2008:83-97；石川，2012:119-124；赤野・堀・投野，2014:210-3 なども同様のカイ二乗検定の方法を紹介している）。この効果量の問題は、効果量の報告自体が一般化していないためか、これま

でそれほど問題視されてこなかった。

第 2 節　単語頻度を使用したカイ二乗検定のケーススタディ

本節では、前節で指摘した統計的有意差と効果量の問題点について、具体的なケーススタディを行って観察する。第 2.1 項では、分析の枠組み、第 2.2 項では分析結果について述べ、第 2.3 項で統計的有意差と効果量の問題点を確認する。

第 2.1 項　分析の枠組み

本項では BCCWJ 出版書籍と白書を使用して、格助詞「が」の使用率の差をカイ二乗検定する分析の枠組みについて述べる。

丸山（2015）は、BCCWJ と『日本語話し言葉コーパス』を使用して、現代語の代表的な格助詞が書籍、新聞、雑誌、白書などの媒体や書き言葉と話し言葉でどのように用いられているのかを観察した研究である。その観察の一つに格助詞「が」の使用率が出版書籍と白書で大きく相違するという興味深い現象がある。先に述べたように、出版書籍は国立国会図書館に所蔵されている書籍のうち、2001 年から 2005 年に発行された約 30 万冊を母集団とし、その中のおよそ 1 万冊から文書を抽出したコーパスである。白書は 1976 年から 2005 年までの 30 年間に発行された 40 タイトル 1,006 冊の白書を母集団とし、そこから 1,500 の文書を抽出したコーパスである。丸山（2015）の観察した現象は、これらの母集団の性質の相違によって生じていると考えられる。そこでここでは「が」の頻度を使用してカイ二乗検定を行い、Cramer's V を算出して効果量の観察を行う。

また出版書籍と白書の効果量を評価するため、母集団の性質が類似していると思われる図書館 SC（以下、図書館書籍と呼ぶ）の効果量も算出し、二つの結果を比較して効果量の問題点を考える。図書館書籍は東京都内にある 13 の自治体の公立図書館で共通所蔵されている書籍のうち 1986 年から 2005 年に発行された約 30 万冊を母集団とし、その中のおよそ 1 万冊から文書を抽出したコーパスである。出版書籍と図書館書籍の主な違いは、発行年の期間と、書籍が出版されたもの全てを含むか（出版書籍）、よく売れている・公共性が高いなどの選択基準で選ばれたものか（図書館書籍）という違いである。

丸山（2015）ではデータに BCCWJ のコアデータ・短単位が使用されている。コアデータとは機械学習用に人手で形態素解析のチェックがなされた、解析精度の高い小規模

データである。ただし統計分析には固定長が適しているとされているため（丸山・柏野, 2014:26；国立国語研究所コーパス開発センター，2015:30）、本研究では固定長・短単位を使用して分析を行う。

第 2.2 項　分析結果

表 6.1 は石川・前田・山崎（編）（2010:60-7）の分析法に従って出版書籍と白書における格助詞「が」の比較を行った分割表である。これをもとにカイ二乗検定を行い、効果量として Cramer's *V* を求めた。これらの値は表の下に示している。表 6.2 は同様の観察を出版書籍と図書館書籍で行った結果である。

表 6.1 「が」の比較：出版書籍と白書

	「が」	「が」以外	総語数
出版書籍	152,010	6,211,425	6,363,435
	2.39%	97.61%	100.00%
白書	16,888	1,024,671	1,041,559
	1.62%	98.38%	100.00%
合計	168,898	7,236,096	7,404,994
	2.28%	97.72%	100.00%

χ^2 =2364.841　p =.000　Cramer's V =.018

表 6.2 「が」の比較：出版書籍と図書館書籍

	「が」	「が」以外	総語数
出版書籍	152,010	6,211,425	6,363,435
	2.39%	97.61%	100.00%
図書館書籍	161,361	6,510,818	6,672,179
	2.42%	97.58%	100.00%
合計	313,371	12,722,243	13,035,614
	2.40%	97.60%	100.00%

χ^2 =12.172　p =.001　Cramer's V =.001

なお固定長は、基本的に約 1,000 字の文書を集積したデータではあるが、第 4 章で観察したように、図書館書籍の場合、4,000 字を超える文書が二つ含まれている。これらにおける「が」の頻度は 126 語と 47 語と多くなるため、表 6.2 では外れ値と認定して分析から除いている。表 6.1、表 6.2 では、p 値がそれぞれ.000 と.001 になり、ともに 5%水準で有意であった。Cramer's V はそれぞれ.018 と.001 であり、APA の効果量の目安で「小」に当たる.10 を大きく下回った。

第 2.3 項　有意差と効果量の問題点

はじめに、表 6.1、表 6.2 における効果量と統計的な有意差の問題点を確認する。表 6.1 の出版書籍と白書における使用率の違いは 0.77%である。使用率の差では互いの相違が分かりにくいためこれを割合に直すと、白書では出版書籍の 67.8%の使用率となっている。これは言語学的に見ても意味のある差だと思われる。しかし Cramer's *V* は.018 であり、APA の目安からすれば実質的な差はほとんどないことになる。表 6.2 における出版書籍と図書館書籍における使用率の差は 0.03%である。これも割合に直すと図書館

書籍では出版書籍の 101.3%の「が」が使用されている。この二つの使用率の差は、言語学的に意味のある差とは思われない。しかし、p 値は.001 で有意になっている。大規模データを使用すると p 値が有意であるからといって言語学的に意味のある差が存在するわけではないことが確認できる。

大規模データで p 値が有意になりやすいという問題は、基本的には言語学的に意味のある差の評価に効果量を使用することで解決できる。しかし単語の使用率の場合、効果量を算出しても今度はその効果量の評価が難しいという問題が生じる。この効果量の問題を解決する方法の一つは APA の目安に替わるコーパス言語学独自の目安を作っていくことであろう。しかしこれには多くの検証と長い時間が必要とされるものと思われる。

より実際的な解決策の一つは、分割表の作り方を変更することである。表 6.1、表 6.2 の分割表では総語数は既知であり、「が」以外の頻度は「総語数－「が」の頻度」で求められるため、新情報は媒体ごとの「が」の合計頻度しかない。合計頻度による比較の場合、媒体間に一定の使用率の違いがあることは分かるが、使用率がどのように異なるかの詳細は分からない。どのように異なるかを詳細に観察する方法については、第 4 節で詳述するが、「が」を何回使用する文書が何文書あるかといった、合計頻度の違いを生み出す要因をもとに分割表の列を細かく区分し、その区分ごとに頻度を比較すればよい。このような使用実態のきめ細かな情報が含まれた分割表なら、媒体間の違いが明確になり、効果量ももっと大きな値を示すと考えられる。APA の目安は心理学の分野で人間を観察単位として行った研究を土台に作成されているため、コーパス言語学でも観察単位を人間に基づいた単位にすることで APA の目安がそのまま適用できる可能性がある。

第 3 節　言語分析における独立性の考察

本節ではコーパスデータにおける独立性の問題を考察する。言語単位が独立性の仮定に違反していることは、すでに第 2 章で全体的に論じている。ここではカイ二乗検定という具体的な分析法において、独立性の問題がどのように影響するかを細かく述べる。第 3.1 項では、カイ二乗検定の原理に照らすと、観察単位の独立性の問題がどのように関係するのかを考察する。第 3.2 項では、単語という言語単位が、どれほど文書に従属しているかを、具体的な用例を通して観察する

第 3.1 項　コーパスにおける観察単位の独立性

はじめに、カイ二乗検定とは、そもそも何を行っているのかを確認する。第 2 節の表 6.1、表 6.2 で使用されているのは、格助詞「が」の頻度と、コーパスに含まれているそれ以外の全ての単語の頻度である。これはコイン投げで表か裏が出る確率などを検定する二項検定の枠組みと同じである。「が」の確率は 2%程度なので年末の福引で回す抽選機をイメージすると分かりやすい。たとえば抽選機の中に赤玉が 2 個、白玉が 98 個入っているイメージである。抽選機に入っている球が合計 100 個なら簡単に当選確率が分かるが、出版書籍の母集団は約 30 万冊の書籍に含まれている言語データであるため、網羅的に調べるのはほぼ不可能である。そこでこの巨大な母集団の抽選機を 6,363,435 回回し、赤玉である「が」が何個出るか調べたとイメージしてみよう。その結果、赤玉は 152,010 回出現したという結果が表 6.1、表 6.2 の出版書籍における格助詞「が」の頻度である。この時点で出版書籍の母集団の当選確率＝「が」の使用率は 2.39%と推定されるが、母集団の巨大さから考えるとまだまだ不正確かも知れない。そこでこのまま最後まで抽選機を回し続けたら、使用率はいったいどれぐらいの値になるのかと考えるのが母数の推定である。この母数の推定を白書でも行った上で、一定の幅を持たせた推定量に差があるかないかを調べているのが検定といえる。正確には二項分布を使用して母数を比較するが、実用的にはデータ数×出現確率が 5 を上回る場合、カイ二乗値を使用しても近似できることが知られている（尾畑，2014:158）。計算はカイ二乗値を算出する方が簡単なのでカイ二乗検定が広く行われてきたという経緯がある。

このカイ二乗検定と観察単位の独立性の問題は、コーパスにおける単語の集め方が、抽選機を回して球を取り出すように、1 個 1 個独立してランダムに抽出されているのかという点にある。BCCWJ の場合、書籍から一定の長さのテキストを切り取った文書を抽出しており、単語を一つ一つ集めたわけではない。「が」は文の中で、名詞類の後に接続して使用されるのが普通である。その文も一文一文が独立しているのではなく、文書という単位に従属して存在している。30 万冊の書籍を集めて抽選機を回せば、それぞれの単語が均一に混じり合うかといえばそうではないだろう。単語同士は数珠つなぎになっていて、均一に混じり合うことはない。抽選機を回して球を 1 個取り出そうとしたら、ずるずるとつながって一まとまりの球が出てしまうようなものである。このまとまりがどこまで続いているのかといえば、母集団であれば 1 冊の本の単位、BCCWJ では、本から抽出された文書の単位と考えられる。

統計的に独立であるとは任意の i 番目と j 番目のデータに関して「j 番目の分布が i 番

目の値に影響されない」ということである（豊田（編著），2009:26）。抽選機の球を単語にした場合、ある単語が選ばれると、単語は数珠つなぎになっていると考えられるため、次に選ばれる単語の選ばれ方に影響を与える。もし、前の単語が後ろの単語に何の影響も与えないのであれば、人間はランダムに単語を選んでいることになるがそれはあり得ない。一方、球を文書にした場合、ある文書が選ばれても次の文書の選ばれ方には何の影響も与えない。コーパスで独立しているのは単語ではなく文書だと考えられる。

第 3.2 項　単語の従属性と文書の独立性の観察

次に、単語という言語単位が、いかに文書に従属しているかを具体的に観察する。単語頻度を使用したカイ二乗検定では、単語が独立した個体であることを前提としている。しかしコーパスにおいて独立した個体は文書であり、単語は文書に従属した存在だと考えられる。ここでは具体的な文書の観察を通し、この考え方の妥当性を検証する。

次の(1)、(2)は新聞の固定長で 1 回しか「が」が出現しない文書の一部である。

(1)　講習会◆ワンデーレッスン「はじめてのトレーニング」　二十四日前十―十一時半、名古屋・池下の愛知厚生年金会館アスレチックルーム。トレーニングの経験のない中高年を対象に、安全で効果的なトレーニングをするためのアドバイスとストレッチの体験。参加費八百円。問い合わせ、申し込み、電０５２（７６１）４１８４。二十三日締め切り。

（PN1f_00001，中日新聞，2001/10/18，夕刊）

(2)　鳴門・堂浦　４号イカダでクモガニ、シラサエビの紀州釣り、二十―四十センチのチヌを十三匹。十一号イカダで小アジを使った泳がせ釣りで三十センチ前後のヒラメを３匹。（戸田渡船０８８・６８８・００５０）

（PN1k_00022，神戸新聞，2001/8/23，夕刊）

(1)、(2)はともに短い字数制限がある場合、極力情報を詰め込もうとする時などに見られる文体で、情報が羅列されている。このような文体の場合、「が」の使用が少なくなるのも当然のように思われる。しかし、(2)と同様に釣りの情報を記述しながら、「が」が頻出する文書もある。

(3)　キスの投げ釣りは十五センチ級が上がる。島見浜南浜漁港、網代突堤で十匹以上。深場は型も良く、メゴチが交じる。ポイントを移動しながらの拾い釣りで三十匹以上釣った人もいた。　　（PN1i_00004，新潟日報，2001/7/6，朝刊）

(3)は固定長で 39 回「が」が出現する文書の一部で、新聞で最も頻度が高い文書の一つである。(2)と(3)では同じ釣り情報を述べながら、ほとんど「が」が出現しない文書と「が」が頻出する文書になっている。(2)では「チヌを十三匹（釣った人がいる）」などの述語が省略されている。しかし「チヌが十三匹（釣れている）」としてもよかったはずである。「チヌが十三匹、ヒラメが 3 匹。」もごく自然な日本語の表現である。もし「を」の替わりに「が」を使用していたら、(2)は「が」が頻出する文書になったと思われる。

(2)と(3)で「が」の出現が分かれたのは、記事によって執筆者の文体の選択が異なったためである。しかも同じ記事が続いている間は、同一の文体が保たれている。(2)で「チヌを十三匹、ヒラメが 3 匹。」としたら、そのつど読者は省略されている述語を変えて記事を読まなければならず、読みにくい。文体そのものの表現効果を売り物にしている文学ジャンルならまだしも、短い文字数の中で多くの情報を伝えようとしている新聞記事の中で、わざわざ読みにくい表現を使用することは考えにくい。(2)、(3)の例を見ると、「が」が文書に従属して出現していることがよく分かる。

「が」には「あの人が田中さんです」のように「あの人」に焦点を当てる際に使用される総記の「が」や、「雨が降ってきた」のように現象文に使用される「が」など必然的に使用される場合もあるが、表現の仕方によっては「が」を使用しないまま文意を伝えることも可能である。最終的に「が」を使うか使わないかは、その文書を記述した執筆者による文体の選択に委ねられている。ただし、文体の選択はどのような媒体にどのような文書を書くか、文字数はどれぐらい与えられているかなど、執筆の目的やさまざまな執筆条件で変化する。また、コーパスの場合、ある文書の執筆者が誰なのか、特定できない場合も少なくない。このためたとえば「が」の使用率を統計的に分析するなら、あるコーパスには「が」が少ない文書や「が」が多い文書がどれぐらい存在するかのように、文書を単位として観察するのが妥当だと思われる。

このような単語と文書の従属関係は、他の研究分野でも確認されている。テキストマイニングの分野では、助詞や助詞相当句を使用した著者推定やジャンルの分類が行われている（金，2002；村田，2000 など）。どのような助詞をどれぐらい使用するかで誰が

書いた文書なのかやどのジャンルに分類される文書なのかが推定できるのは、助詞の頻度が文書（執筆者）に従属しているからだと考えられる。

第 4 節　文書度数分布の観察と効果量の確認

本節では文書を個体、「が」の頻度を変量とした文書度数折れ線を描き、出版書籍、白書、図書館書籍における文書度数分布の違いを観察する。また、文書を観察単位としてカイ二乗検定を行うと、効果量の評価がしやすいことを述べる。第 4.1 項では出版書籍の文書度数折れ線を描き、頻度と文書の特徴について観察する。第 4.2 項では白書と出版書籍の文書度数折れ線を比較し、第 4.3 項で白書の「が」の使用率が低い理由を考察する。第 4.4 項では図書館書籍の文書度数折れ線が図書館書籍と酷似している状態を観察する。第 4.5 項では、文書を観察単位とした出版書籍と白書、出版書籍と図書館書籍のカイ二乗検定を行って効果量を算出し、文書を観察単位とすれば効果量の評価が行いやすいことを述べる。

第 4.1 項　出版書籍における文書度数分布の観察

はじめに、出版書籍の文書度数折れ線を描き、この図を利用して文書度数分布の特徴を観察する。図 6.1 は出版書籍の度数折れ線である。横軸は固定長・短単位の文書に出現した「が」の個別調整頻度、縦軸はその「が」の頻度が出現する文書の度数を表している。

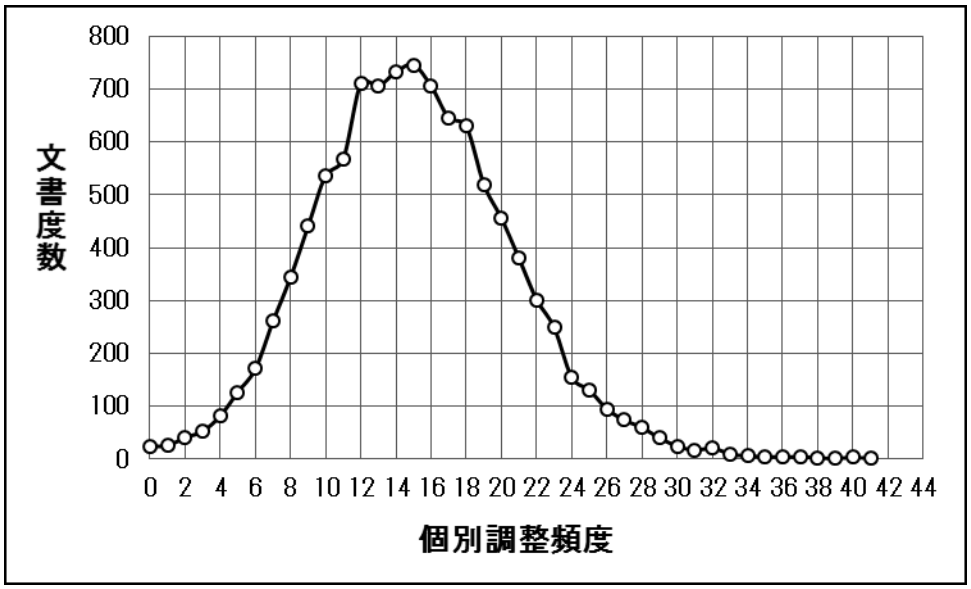


図 6.1 格助詞「が」の文書度数折れ線：出版書籍

固定長・短単位の語数は平均 631.4 語、標準偏差 59.5 語とばらつきがあるため、すべての文書を 631 語に調整した頻度で描いている。個別調整頻度＝「が」の粗頻度÷文書の語数×631 語である。図 6.1 で横軸が 0、つまり 631 語当たり 1 回も「が」が使われていない文書は 10,117 文書中 25 文書ある。これは「が」の使用率 0%の文書が 25 文書あるという意味と同じである。図 6.1 の頂点に当たる頻度 15 の使用率は 15÷631 語＝2.38%なので、これが表 6.1 の使用率 2.39%とほぼ同じになる。頻度 32 の使用率は 32÷631 語＝5.07%で、「が」の使用率が 5%を超える文書はほとんど存在しないことが分かる。図 6.1 は、「が」をどれぐらい使用している文書がどのように分布しているかを表しているグラフだが、横軸を頻度ではなく使用率に読み替えると、文書ごとの使用率の分布を表しているグラフだと見なすことができる。出版書籍は、サンプルサイズが 10,117 あり、ジャンルの多様性も十分であるためか、文書度数折れ線は正規分布に近い曲線を描く。「が」の中央値は 15、平均が 15.03、標準偏差は 5.65 である。

「が」はかなり狭い範囲の頻度で使用されており、15±5 回の区間で使用されている文書だけで全文書の 68.8%になる。15±5 回という頻度は、「が」の使用が多いのか少ないのか、一読しても分からないレベルである。以下に用例を示す。(5)は「が」の頻度が 10、(6)は 15、(7)は 20 の文書の一部である。

(5) 現代の日本社会において高齢化という言葉は、われわれの日常生活のなかにすっかり浸透してしまった感がある。いまや高齢者世代は、他の世代に比べて最も注目をされる存在となっている。

（PB23_00343，安達正嗣，『高齢化と少子社会』）

(6) 知的所有権、特に特許は、もともと創造された新技術を世界に公開するための仕組みとして生まれた。すなわち新技術を発明したとき、その技術を発明者がこっそり使うのでは人類全体の進歩に貢献しない。そこでその発明を公開して、誰でも使えるように決めたのである。

（PB43_00764，桑原裕，『技術経営とは何か』）

(7) 「企業会計原則」によって財務諸表体系の中心に損益計算書が置かれるまで、わが国の会計学の中心的な研究対象は貸借対照表であった。したがって貸借対照表論が会計学の代名詞だったのである。これは、わが国の会計学がドイツ会

計学の強い影響のもとにあったからである。

（PB33_00477，中村忠,『新稿現代会計学』）

(5)〜(7)は一部分の引用であるため、文書全体での「が」の出現状況は比較できないものの、それぞれの文体的特徴は確認できる。とはいっても、どれも一般的な説明文で、これらの文書で何が「が」の頻度の違いを生み出しているのかよく分からない。それにも関わらず、図 6.1 を見ると、15 回の平均から離れるにつれて文書数は着実に減少するため、「が」の分布には意識では捉えにくい産出のルールが確実に作用していると思われる。

一方、頻度が極端な文書の場合は、文体的特徴がもう少し分かりやすい。図 6.1 は正規分布に近いが、右側の裾が厚く、数は少ないものの一部に「が」を多用する文書があることが分かる。出版書籍で「が」の出現が最も多かったのは次の(8)の文書である（頻度 44）。

(8) 強迫傾向や被害・関係念慮に関する領域　表７に日本と中国の大学１・２年生の各質問項目におけるチェック率を示した。この領域全１０項目の平均チェック率は、日本男子１年が２１．７％、同２年が２４．７％、日本女子１年が２８．７％、同２年が２８．９％、中国男子１年が２７．５％、同２年が２５．０％、中国女子１年が３０．６％、同２年が３０．４％であった。

（PB43_00201，白佐俊憲ほか（編著）,『日本と中国の大学生の精神的健康』）

(8)は表の内容を細かく説明する記述になっており、このように一つ一つの項目について数値を列挙して述べる文書では、「が」が頻出するのも理解できる。これと反対に「が」の頻度が 0 の文書には、(9)、(10)のように料理の作り方やソフトの使い方など、手順を示す文書が見られる。これらの文書では「何を、どうするか」の説明が中心となるため、「何が」「誰が」に当たるものが記述されない。その結果、「が」が出現しないと考えられる。

(9) 散らし寿司《材料》《調理方法》１　にんじんともどした干ししいたけはみじん切りにし、だし汁、砂糖、みりんで煮、しょうゆで味つけをし、汁気をきっておきます。

（PB14_00069，藤本真美子（編著）,『これで完璧！！高齢者の食事ケア』）

(10) フォントの色は［ミキサー］パネルで灰色（ＲＧＢをそれぞれ［１０２］）にし、線の太さは［線］パネルで「極細線」とします。

（PB17_00136, 実著者不明,『Flash web design book ver.5j』）

「が」の頻度が極端な文書の多くは(8)～(10)のような特異な文体を持っている。出版書籍では「が」が少ない文書と「が」が頻出する文書に挟まれて、「が」が平均的に出現する文書が大量に存在し、「が」の分布はほぼ正規分布に従うと考えられる。

第 4.2 項　白書における文書度数分布の観察

次に、白書における「が」の文書度数分布を観察する。図 6.2 は、白書と出版書籍の文書度数折れ線を並べて描いたグラフで、図 6.1 の出版書籍との比較を可能とするため、相対度数折れ線で描いている。縦軸は文書度数の比であるから、白書における文書の絶対度数を求めたい場合は、この比にサンプルサイズの 1,500 をかければよい。また出版書籍の固定長・短単位の平均語数は 631.4 語、白書は 694.6 語と 1 割ほど異なるため、何も調整しない生データで比較すると白書の「が」が 1 割ほど多くなってしまう。このため出版書籍、白書ともに図 6.1 と同様に 631 語当たりの個別調整頻度で描いている。

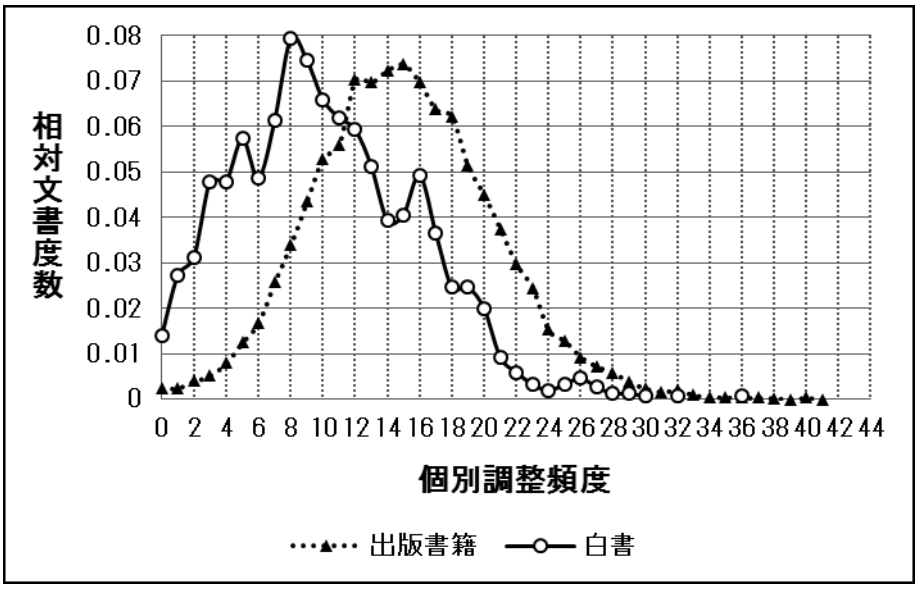


図 6.2 格助詞「が」の相対文書度数折れ線：出版書籍・白書

図 6.2 の白書における「が」の文書度数分布は、出版書籍に比べ左側に寄っている。これは白書において、「が」の頻度が低い文書が多いことを表している。「が」の中央値

は 10、平均が 10.22、標準偏差は 5.66 である。

なお、出版書籍が滑らかな曲線であるのに対し、白書の線が上下に揺れるのは、白書のサンプルサイズが 1,500 と小さいためだと考えられる。第 1 節ではデータ数が多いと多くの場合で *p* 値が有意になることを述べた。このためデータ数が多すぎない方が分析に適しているように思われやすいが、文書度数分布を観察する場合には 1,500 のデータ数でも十分ではないと考えられる。たとえば BCCWJ のコアデータの場合、形態素解析の精度は高いが出版書籍の文書数は 83 に過ぎない。この場合、解析精度の高さというメリットよりも統計的な誤差のデメリットの方が大きく、これで図 6.2 のような度数折れ線を描くと、統計的な代表性を持った図 6.2 の分布からはかけ離れた図となる。比較したコーパスに実質的な差があるかないかは *p* 値ではなく効果量を算出して判断すればよいため、基本的にデータ数は多い方が望ましいと考えられる。

第 4.3 項　白書で「が」の使用率が低い理由

白書で「が」の使用率が低い理由は何だろうか。本項ではこの問題を考察する。図 6.2 のような文書度数分布図を描くと、出版書籍と白書における「が」の使われ方の違いが詳細に観察できる。さらにこの分布図の元データを利用すると、「が」を使用していない文書や逆に多用している文書などを簡単に特定することができる。

ここで白書における特徴的な用例を観察してみよう。(11)〜(13)は、「が」の頻度が 0 の文書である。

(11) ⅱ　家計収支及び物価の動向　全国勤労者世帯の実収入を見ると、前年同期比（実質）で、９年１～３月期３．１％増、４～６月期０．４％増、７～９月期２．８％増、１０～１２月期１．３％減となり、年間の伸び率は１．１％増と前年の１．５％増に比べ低下した。消費者物価は安定的に推移し、９年の消費者物価（持家の帰属家賃を除く総合）の対前年上昇率は１．６％となった。

（OW5X_00196,『観光白書』, 平成 10 年度版）

(11)は、出版書籍の例文(8)とよく似ており、グラフや表などの各項目ごとに値を記述する文体である。(8)は、「が」の頻度が 44 と出版書籍の中で最も「が」が多い文書であった。(11)も「９年１～３月期**が**３。１％増」のように「が」を入れると頻度が極端に多くなる。しかし、白書では冗長性を嫌うためか、このような場合「が」を入れない

文体で書かれた文書が多い。この結果、「X は Y○○%」のような文体が続く報告となり、「が」が出現しない。

(12) 大蔵省においては、地方公共団体に対する資金運用部資金の貸付予定額を次のとおり決定した。地方長期資金等の貸付（３）　国民金融公庫の融資　国民金融公庫においては、被災中小企業者に対し、次のとおり災害融資を行った。（OW4X_00491,『防災白書』, 平成 5 年版）

(12)は、事業主体を「～においては」で示し、事業内容を列挙する文体である。「～においては」で主体が明示されるため、「が」で主格を示す必要性がない。白書の主目的は国会に対して事業の実施や計画を報告することにあるため、その事業内容の報告が続いている間は、(12)のような文体が繰り返されることになる。

(13) ３　航空機の検査体制の充実　航空機の大型化、高性能化に対応した検査体制の充実を図るため、航空機検査官の研修の充実を図り、その質的向上に努めた。４　航空機の整備審査指導体制の充実　航空機の整備に対する審査及び指導体制の充実を図るため、〔･･･〕我が国での航空機の使用形態に適合した適切な整備方式を確立するよう航空会社を指導した。（OW3X_00119,『交通安全白書』, 昭和 63 年版）

(13)も(12)と同様に実施した施策を報告する文体である。この場合、事業主体が同一であるため、逐一「～においては」で示されることはない。ここでも実施した事業内容が続いている間は同じような文体が繰り返され、その間、「が」が出現することはまれである。

ただし、項目列挙型の文書でも、単に実施した内容を列挙するのではなく、実態の報告を分析的に述べる場合には、「が」が多用されることもある。(14)は、防災訓練の課題について述べた文書である。この文書全体では「が」が 22 回出現する。

(14) 住民の参加に関する課題　一般住民の防災への取組みは、災害の体験や災害に関する報道の有無によって左右されることが多く、このため防災訓練に常に多くの参加を得ることは困難な面もある。特に、訓練が平日に行われる場合、勤め人は訓練に参加しにくいため参加者が固定化する傾向が見られるほか、一般住民が避難訓練を行っている傍らで、平常業務を続けている事業所等がある

など緊張感のある訓練ができないなどの問題もある。

（OW3X_00073,『防災白書』, 昭和 61 年版）

同じように概念の説明や執筆者の考えを述べる文書では、「が」が多用される傾向にある。特に施策の実施ではなく、内外の経済動向に関する調査・分析を主要な業務とする経済企画庁が発行する経済白書ではこの傾向が強い。(15)は、経済白書で「が」が 32 回使用されている文書の一部である。

(15) 資産価格はファンダメンタルズだけで決まるものではない。長期金利や収益があまり変動していないにもかかわらず、資産価格が短期間に大幅に変動することがある。そのような資産価格の大幅な変動のなかには、収益の増加率または資産価格の上昇率の期待値に大きな変化が生じたことによると考えられる場合もある。（OW4X_00348,『経済白書』, 平成 3 年版）

以上の例から分かることは、白書という媒体が国会への報告書という資料的性格を強く持ち、事業の実施や計画を報告する白書特有の文体を使用しているため、「が」の使用が少なくなるという傾向である。逆に概念の説明や執筆者の考えを述べる文書では、「が」が多用される場合もある。最終的にはどのような性格を持った文書が、白書全体でどのように分布しているかによって、「が」の使用率が決まる。

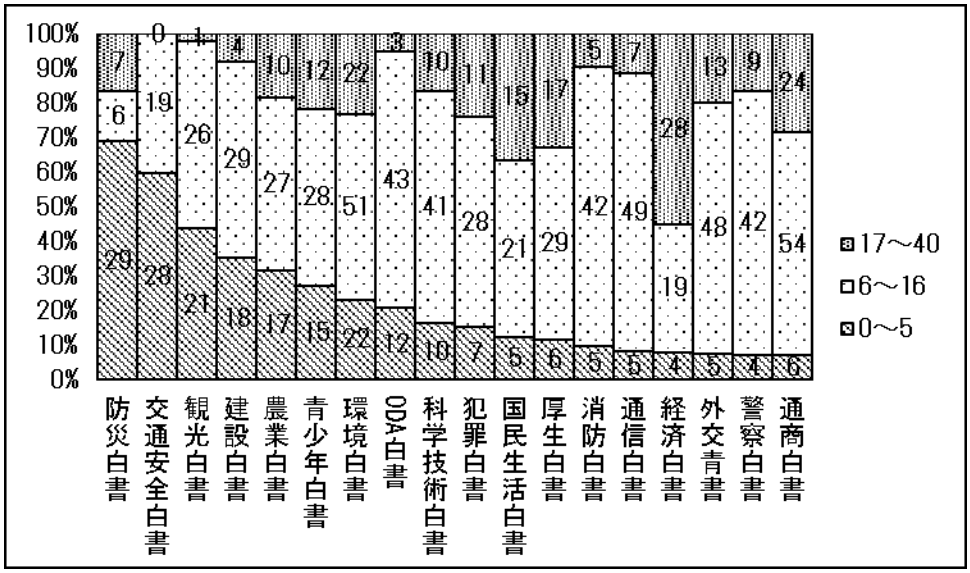


図 6.3 白書の主要ジャンルにおける「が」の頻度割合

図 6.3 は文書数が 41 以上の主要な白書について、「が」の頻度を少なめ（0~5）、中間（6~16）、多め（17~40）の三つに区分してその割合を描いたグラフである。図 6.3 では事業の実施報告を主目的とする防災白書や交通安全白書では、少なめの文書が 60%以上を占めている。一方、現状の調査や分析を主目的とする経済白書や国民生活白書などでは、「が」が多めの文書が多い。しかし、白書全体から見た場合、頻度が 17 を超えるようなジャンルの白書は少ない。これは白書のそもそもの性格が、行政が行った施策の実施報告にあるためだろう。このような理由によって、図 6.2 の文書度数分布図における白書は「が」が少ない文書が多く、「が」を多用する文書はその頻度が大きくばらついて右裾広がりの分布となっていると考えられる。

第 4.4 項　図書館書籍における文書度数分布の観察

次に、出版書籍と図書館書籍の分布を比較して観察する。図 6.4 も出版書籍と比較するため、相対度数折れ線と 631 語当たりの個別調整頻度で描いている[21]。これを見ると図書館書籍の分布は、ほぼ出版書籍と同じ分布をしていることが分かる。統計量は出版書籍、図書館書籍の順に中央値が 15、15、平均が 15.04、15.24、標準偏差が 5.58、5.36 である。

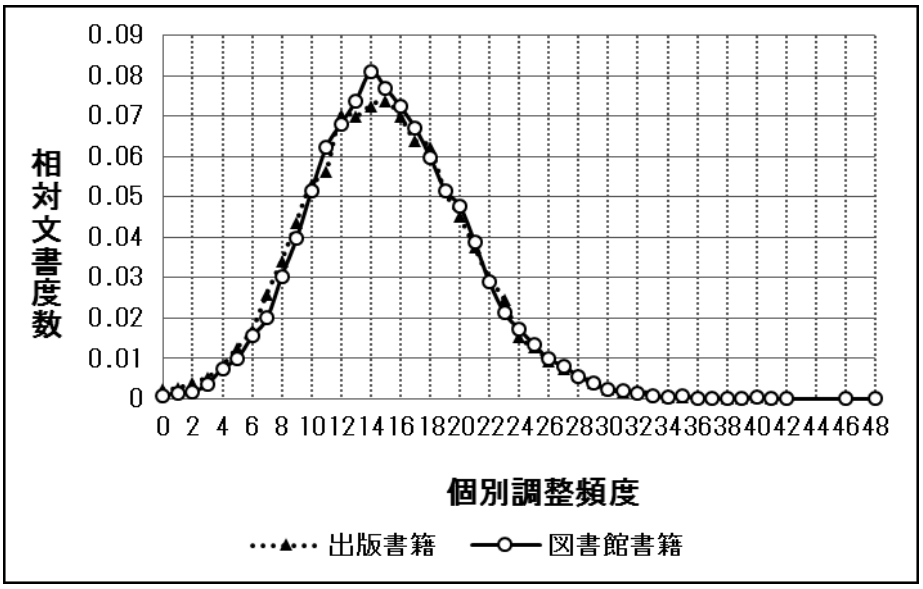


図 6.4 格助詞「が」の相対文書度数折れ線：出版書籍・図書館書籍

[21] 表 6.2 の分割表では語数が多い文書を外れ値として除いたが、ここでは相対化した頻度を使用するためこれらも分析に含めた。相対化後の頻度は 126 語→17 語，47 語→12 語となる。

出版書籍と図書館書籍の主な違いは発行年の期間と、書籍が出版されたもの全てを含むか（出版書籍）、よく売れている・公共性が高いなどの選択基準で選ばれたものか（図書館書籍）という点にあるが、図 6.4 の格助詞「が」の分布ではほぼ同じ性質を持っていることが観察できる。

第 4.5 項　文書度数を使用したカイ二乗検定と効果量の観察

最後に、図 6.2、図 6.4 で観察した文書度数分布の差が母集団に一般化して成立するかどうかを、カイ二乗検定を使用して検定するとともに、クラメールの連関係数を算出して効果量の大きさを確認する。表 6.3、表 6.4 は、「が」の頻度を 6 個単位で区切り、その区間に落ちた文書度数で分割表を作っている。

表 6.3 出版書籍と白書における格助詞「が」の文書度数比較

	0～5	6～11	12～17	18～23	24以上	合計
出版書籍	354	2,322	4,248	2,537	656	10,117
	3.50%	22.95%	41.99%	25.08%	6.48%	100.00%
白書	339	588	415	132	26	1,500
	22.60%	39.20%	27.67%	8.80%	1.73%	100.00%
合計	693	2,910	4,663	2,669	682	11,617
	5.97%	25.05%	40.14%	22.97%	5.87%	100.00%

χ^2 =1204.243　p =.000　Cramer's V =.322

表 6.4 出版書籍と図書館書籍における格助詞「が」の文書度数比較

	0～5	6～11	12～17	18～23	24～29	30以上	合計
出版書籍	354	2,322	4,248	2,537	555	101	10,117
	3.50%	22.95%	41.99%	25.08%	5.49%	1.00%	100.00%
図書館書籍	269	2,318	4,634	2,620	614	96	10,551
	2.55%	21.97%	43.92%	24.83%	5.82%	0.91%	100.00%
合計	623	4,640	8,882	5,157	1,169	197	20,668
	3.01%	22.45%	42.97%	24.95%	5.66%	0.95%	100.00%

χ^2 =23.713　p =.000　Cramer's V =.034

カイ二乗検定ではセルの期待度数が 5 未満になると精度が下がるため、表 6.3 では「が」の頻度 24 以上、表 6.4 では「が」の頻度 30 以上の文書度数は集約した。この表をもとにカイ二乗検定を行ったところ p 値はいずれも.000 になり 5%水準で有意であった。図 6.4 では出版書籍と図書館書籍には、実質的な差はほとんど差が見られないが、表 6.4 の p 値は有意となる。これは観察単位を単語から文書に変えても個体数が約 2 万と非常に大きいためである。

一方、クラメールの連関係数は表 6.3 が.322 で、APA の目安に従えば出版書籍と白書の分布は中程度に異なっていると評価できる。一方、表 6.4 の値は.034 とごく小さく、出版書籍と図書館書籍の分布はほぼ同じであるといえる。これらの値は図 6.2、図 6.4 のグラフの重なりの程度を数値化した値である。この分析法では連関係数は二つのグラフが完全に重なれば 0、重なりが全くなければ 1 となるため、直観的にも分かりやすい。

文書度数分布図はどのような文体を持った執筆者がどれぐらいいるかを表している分布図と考えることもできる。これは質問紙調査や実験で人間を個体として観察しているのと同じ枠組みであるから、連関係数はコーパス分析で得られた値でも、質問紙調査や実験などで得られた値でも同じ基準で評価できることになる。

第 2 節で行った単語頻度を使用したカイ二乗検定では、コーパスに含まれている単語を独立した個体として見なすという分析の枠組みそのものに疑問があり、使用率という小さな値の差をもとに算出された効果量も評価が困難であった。一方、文書度数を使用したカイ二乗検定では、分析の枠組みも明解で、得られた効果量も評価しやすい。サブコーパス間における格助詞「が」の使用率の相違は、文書度数分布の相違で説明ができる。よってコーパス間における使用率を比較する場合、文書を観察単位とする分析法が有効だと考えられる。

第 5 節　まとめ

本章の目的は先行研究で指摘されているカイ二乗検定の問題点を再考し、統計学的にも言語学的にも有効なカイ二乗検定の方法を提案することにあった。先行研究で指摘されている問題点の一つは大規模データを使用すると実質的に意味のない差であっても p 値が有意になるという問題である。この問題は差の評価を p 値ではなく、効果量に替えることで基本的には解決できる。ただし単語の使用率のようにごくわずかな差を分析する場合、クラメールの連関係数の評価に APA の目安は使えない。APA の目安は人間を観察単位とした心理学の分析をもとに作られているため、使用率の比較でも執筆者の意図や個性が反映されている文書を観察単位とし、互いの差が詳細に観察できる分割表を作成する分析法が有効だと考えられる。

先行研究で指摘されている問題の 2 点目は、言語データには統計的検定の前提となる独立性が欠如しているという問題である。コーパスにおいて独立した個体は単語ではなく、文書だと考えられる。このため独立性の観点からも文書を観察単位とした分析が有効であると考えられる。観察単位とは、データを収集した時に何を個体と見なして観測

値を集めたかというデータ収集にかかわる単位である。従来の研究では、文字、単語、文などの言語単位が無作為抽出されていると考えてきたため、これらを観察単位としてカイ二乗検定を行ってきた。しかし、現実的に無作為抽出されているのは文書であり、文書内の文字、単語、文は互いに強い関連性を持っていて独立していない。このため、文書を観察単位にして統計分析を行う必要がある。

文書を観察単位として文書度数分布図を描くと、単語の使われ方の特徴が詳細に観察でき、元データで頻度に特徴のある文書を特定するのも容易である。また文書度数を使用してカイ二乗検定を行い、クラメールの連関係数を求めると、APA の目安を適用して効果量を有効に評価することができる。このためコーパス間で単語の使用率の比較を行うなら、単語ではなく文書を観察単位とすべきだと考えられる。文書を観察単位とするカイ二乗検定の方法を一般化してまとめると以下のようになる。

〈文書を観察単位とするカイ二乗検定の方法〉

1. コーパスを構成する独立した個体を文書と考える。
2. 単語の頻度は文書ごとに変化する変数と考える。
3. 文書の語数で標準化した変数（個別調整頻度）を一定の間隔に分割し、その区間（階級）に落ちた文書度数を求める。
4. 文書度数で分割表を作り、カイ二乗検定を行って効果量を求める。

分割表では、カテゴリーごとに度数を集約するが、この度数は個体の数であって、変数ではない。分割表は、あるカテゴリーに分類される個体がいくつあるかを記した表であり、個体の観測値（変数）の合計数を記しているのではない。これに対し、平均値や中央値というのは観測値（変数）の要約値であり、度数ではない。この度数と変数の関係を混同しないことが重要である。

コーパス言語学では、長らく単語の数を頻度と呼んできた。これは単語を個体と考えてきたためである。第 3 章で記したように、本研究では単語の数は変数と考えるため、本来これを頻度と呼ぶことはふさわしくないが、コーパス言語学の慣習に従って頻度という呼び方を踏襲している。このため、「2. 単語の頻度は文書ごとに変化する変数と考える」という表現では、「単語の観測値」のように記述する方がより適切である。

第 7 章　回帰分析の方法

本章では、コーパスを使用して統計学的にも言語学的にも有効な回帰分析を行う方法を検討する。回帰分析は、調査対象の相関関係に基づいて因果関係の解明を行う最も基本的な分析法の一つであり、コーパスを使用した言語分析においても積極的に活用することが望まれる。しかし、単語などの言語単位を観察単位とした場合、一つ一つの単語に頻度情報としての変数は存在しないため、個体レベルの回帰分析を行うことは難しい。森（2011:50）では、相関分析の解説の中で「コーパス調査で相関を調べるというのは難しいため、アンケート調査か言語使用調査のデータを扱うことになる」と述べられている。このため、コーパスを使用して相関関係を観察する場合は、書籍や新聞などの出版媒体や、文学や哲学などのジャンルのように何らかのカテゴリーで頻度を集約し、集団レベルの相関係数を求める分析が行われている（石川・前田・山崎（編）, 2010:85-104）。

しかし、本研究の分析法では、個体レベルの回帰分析が可能になる。本章ではこれまでコーパス言語学で行われてきた集団レベルの回帰分析と、本研究が提案する文書を観察単位とした個体レベルの回帰分析の比較を行い、個体レベルの分析を行う重要性を示す。分析には BCCWJ 固定長・長単位データを使用する。

第 1 節では、集団レベルの回帰分析と個体レベルの回帰分析を比較し、これらが本質的に異なる分析であることを確認する。第 2 節では、集団レベルで分析すると、本当は相関がないのに相関があるかのように誤って推論する生態学的誤謬（Robinson，1950；森，1987）や分割相関に配慮しない分析によって誤謬が起きる例を示す。第 3 節～第 5 節では、文書を観察単位とすると、データのかく乱要因を避けるために変数を精緻化したり、文体分析に不向きな文書を除くことが可能になり、精度の高い回帰分析が行えるという個体レベルの分析法のメリットを示す。最後に第 6 節で、本節のまとめを述べる。

第 1 節　集団レベルと個体レベルの回帰分析の違い

本節では、樺島（1955）に準拠して、集団レベルと個体レベルの回帰分析を行い、これらが本質的に異なる研究であることを確認する。第 1.1 項では、先行研究の分析例を通し、集団レベルと個体レベルの回帰分析の違いが認識されにくいことを述べる。第 1.2 項では、樺島（1955）に準拠した分析を行うためのデータについて説明する。第 1.3 項では、分析結果を検討し、集団レベルと個体レベルの回帰分析では、何がどのように異なるのかを観察する。

第 1.1 項　先行研究と分析の目的

回帰分析は、集団レベルで行われる場合と個体レベルで行われる場合があるが、これらの分析レベルの違いは、簡単には認識されにくい。たとえば樺島（1955）と樺島・寿岳（1965:29）は、ともに日本語の品詞構成比率を扱った回帰分析の研究であるが、結論がよく似ているため、これらの研究の違いが簡単には認識されにくい。樺島（1955）は、樺島（1954）の結果に追試を加え、名詞の比率を説明変数、他の品詞群の比率を目的変数として回帰分析を行った研究で、日本語の品詞構成比率には一定の傾向性が存在することを明らかにした。日本語の品詞グループの出現比率を定式化した数式は樺島の法則と呼ばれている。図 7.1 は樺島（1955）のデータを使用して描いた散布図と回帰直（曲）線である。一方、これとよく似た研究に、樺島・寿岳（1965:29）の回帰分析がある。図 7.2 は樺島・寿岳（1965:29）所載の図 2.1 と同様に、付表 A（pp.219-222）に基づいて作成した散布図である[22]。

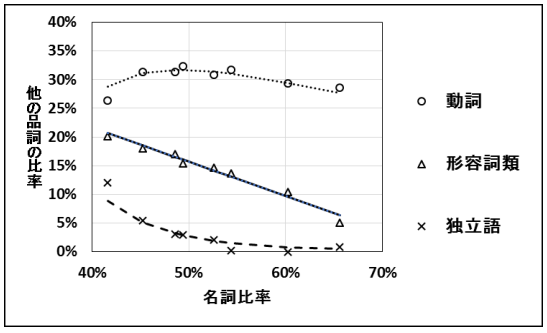


図 7.1 樺島（1955）第一表に基づく散布図

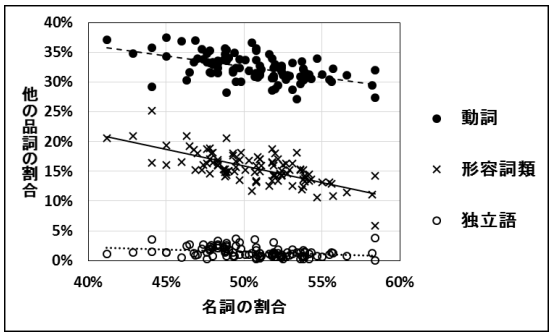


図 7.2 樺島・寿岳（1965:219-22）付表 A に基づく散布図

この二つの研究は、分析レベルが異なり、図 7.1 は集団レベル、図 7.2 は個体レベルの分析を行っている。樺島（1955）では、文章の種類というカテゴリー別に 300 文（和歌・俳句は 118 句）が無作為抽出され、カテゴリー別に計算された品詞比率が回帰分析に使用されている。これは、観察単位が文単位、分析レベルが集団レベルである。図 7.1 の散布図のマーカーは左から順に日常会話、小説会話、哲学書、小説地の文、自然科学書、和歌、俳句、新聞記事の品詞比率となっている。一方、樺島・寿岳（1965:29）は 100 編の小説を観察単位とし、作品別の品詞比率を分析に使用している。これは観察

[22] 樺島・寿岳（1965:29）所載の図 2.1 は、付表 A に対し、ややマーカーの位置にずれが見られる。このため、図 7.2 は、樺島・寿岳（1965:29）所載の図 2.1 とは、若干形状が異なっている。

単位が小説の作品、分析レベルが個体レベルである。

この二つの分析は同じ品詞比率を扱ってはいても、分析の内容も結果も異なる。しかし、これらの論文で主張されている内容はよく似ている。

> 以上のように、條件を一定にした文において、名詞の百分率がわかれば、他の類別された品詞の百分率が算出できること、及び條件が異なる文を比較すると、名詞の百分率に差が見られることから、次の事が考えられる。
>
> 即ち、名詞の百分率をもつて、文章の特性を計る尺度となし得る（後略）
>
> （樺島，1955:386）

> この図からわかるように名詞 N の比率がわかれば，他の品詞の組がどのような比率をもつかはだいたい見当がつく．
>
> また名詞は品詞中もっとも大きな比率をもつから，品詞比率の代表値として名詞の比率を使うことができる．（樺島・寿岳，1965:29）

樺島・寿岳（1965:29-36）では、集団レベルの分析である樺島（1963）[23]の分析結果に基づいて、個別の小説作品の名詞比率が論じられており、集団レベルの分析と個体レベルの分析が特に区別されていない。このためか、先行研究の中には樺島の法則を、個体レベルの法則だと考えるケースが見られる。白井・三浦（2012）は、日本語の文書のジャンル分類に、品詞分布を使用した研究である。この中で、樺島の法則について「ジャンルによる品詞分布の特性を考慮していないため、当てはまりがあまり良くない」と述べられている。白井・三浦（2012）は、文書分類という個体レベルの研究を行っているのに、集団レベルの研究結果である樺島の法則を使用したため、「当てはまりがあまり良くない」という結論になったと思われる。

文書分類とは、文書のジャンル等を機械的に判定し、文書を分類する研究分野である。文書の品詞比率は形態素解析技術によって機械的に求めることが可能だが、その文書が属するジャンルは、基本的には人間が読んで判断するしかない。それを機械的に分類できるようにするのが文書分類の研究目的である。白井・三浦（2012）では、図 7.1 のグラフを使用すれば文書のジャンル判定ができると考えたと思われるが、このような目的であれば、図 7.2 のような個体レベルのグラフを、小説以外のさまざまなジャンルでも

[23] 樺島（1954，1955）と同様の分析をもう一度行った研究である。

作成し、それら複数の回帰式から推定するのが妥当だと思われる。樺島の法則は、ジャンル平均という集団の分布を表しているため、同じジャンルの文書を数十～数百集めてその平均を出した時にはこのグラフに当てはまる可能性はあるが、一つの文書という個体レベルのジャンル判定が行えるような研究の枠組みにはなっていない。

このように、集団レベルの回帰分析と個体レベルの回帰分析は、類似した分析に見えても異なった研究であり、集団レベルの回帰分析を行ったからといって、個体レベルの因果関係が解明できるわけではない。

本節では、集団レベルの回帰分析と個体レベルの回帰分析では、何がどのように異なるのか、なぜ集団レベルの回帰式を使用すると、個体の判定精度が高くならないのかについて、樺島（1955）に準拠した簡易的な調査を行い、その仕組みを観察する。

第 1.2 項　分析データの説明

分析には BCCWJ 図書館書籍の一部と新聞の固定長を使用する。言語単位は樺島（1955）に合わせ、文節を基準として認定された長単位を使用する[24]。樺島（1955）では 1 文単位でデータが集積され、300 文の平均が使用されている。しかし、本節の分析目的は樺島（1955）の完全な追試ではなく、集団レベルと個体レベルの回帰分析の違いを観察することにあるため、これまでと同様に文書を観察単位とした分析を行う。

樺島（1955）では 8 種類のカテゴリーに分けた言語データが使用されているが、本節では「対話」379 文書、「哲学」522 文書、「自然」642 文書、「新聞」1,473 文書の 4 種類で観察を行う。「対話」は、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）[25]の分類に従い、図書館書籍に含まれている文書で対談やインタビュー、シナリオなど基本的に対話とみなせる文書を使用する。国立国語研究所（2015）は、複数の研究者が図書館書籍の 10,551 文書を一つ一つ読んで判断した文体情報を付与したデータで、その詳細は柏野（2013）で紹介されている。哲学書、自然科学書は、図書館書籍に付与された日本十進分類法（NDC）に従って認定し、名称は「哲学」、「自

[24] 山崎（2016:132）は「樺島（1954）（1955）には，品詞の計測に際して文節を使用したとする記載は見当たらない」という慎重な立場を取るが，樺島（1954:15）には「文節数＝自立語数」，「所謂自立語と総称される品詞（名詞・動詞・・・・）」とあり、樺島（2009:95）では樺島の法則の説明に「品詞を，橋本進吉が定義した文節に含まれる自立語を単位として判定し」とあるため、本研究では樺島（1954）とその追試である樺島（1955）は、文節に基づいた分析だと考える。

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

然」と略す。「新聞」は出版 SC 新聞レジスターをそのまま使用する。以下は、(1)「対話」、(2)「哲学」、(3)「自然」、(4)「新聞」の用例である。用例は、それぞれのカテゴリーの平均値に近い品詞比率となっている文書の一部を引用した。

(1) 田辺　そうすると、寛美さんはどんなタイプの女性がお好みなの？寛美　十五、六歳のころ、高峰三枝子さんが好きでした。田辺　ずうっとそのまま？寛美　変形はしてきていますけれども、やっぱり背の高い、やせ型の女性。田辺　日本風の美人？寛美　というより、舌たらずみたいな、甘えたみたいな女性ですね。

（「対話」: LBa9_00011，田辺聖子・藤山寛美,『おせいさんのほろ酔い対談』）

(2) 私の住んでいる田舎町では、サラリーマンと結婚した女性の多くが結婚と同時に家庭に納まり、子供を二人生んで子育てに没頭し、末の子が幼稚園に入る頃になると、母親として活発に社会参加をしたり、またはパートで働きに出たりするようになる。　母親としての社会参加は、まず子供の成長につれて幼稚園の母の会、小学校、中学校のＰＴＡの役員として活躍することである。

（「哲学」: LBn1_00013，松田幸子,『女性を生きるための哲学入門』）

(3) 夫婦ともに、新居に入るまでは、日常生活でのアレルギーなど特別な症状はなかった。なぜこんな異変が起きたのだろうか。　“問題の家”を検証してみよう。　たとえば、玄関の上がりがまちの下側から床下に抜ける部分に縦七センチメートル、横幅百六十センチメートルもの大きな隙間がある。

（「自然」: LBq4_00006、木村元紀・柳沢幸雄、『化学物質過敏症』）

(4) 米誌ニューヨーカー（電子版）は十五日、バグダッド郊外のアブグレイブ刑務所で発覚したイラク人虐待事件に関し、ラムズフェルド米国防長官が昨年、これまでより過酷な尋問方法をイラクで行うよう奨励する「極秘作戦」を承認、これがイラク人虐待につながったと報じた。

（「新聞」: PN4c_00009，読売新聞，2004/5/17，朝刊）

これらの用例で明らかに文体が異なるのは(1)で、(2)と(3)にはそれほど大きな違いは感じられない。(4)は報道文特有の文体で、(2)、(3)とは少し異なる印象である。このよ

うな文体の違いが、品詞比率の違いになって表れると考えられる。

第 1.3 項　分析結果と考察

図 7.3～図 7.6 は「対話」「哲学」「自然」「新聞」の 4 種類のカテゴリーごとに、名詞の比率を説明変数（x 軸）、動詞、形容詞類（形容詞＋形容動詞＋連体詞＋副詞）、独立語（接続詞＋感動詞）の比率を目的変数（y 軸）として散布図を描き、回帰直線を書き入れた図で、個体レベルの分析である。これらの回帰直線 $y = a + bx$ における傾き（b）と切片（a）、並びに回帰直線の当てはまりのよさを示す R^2（決定係数：ピアソンの積率相関係数を 2 乗した値で、説明変数が目的変数の分散の何%を説明しているかを表すため、分散説明力とも呼ばれる）は表 7.1 に示した。

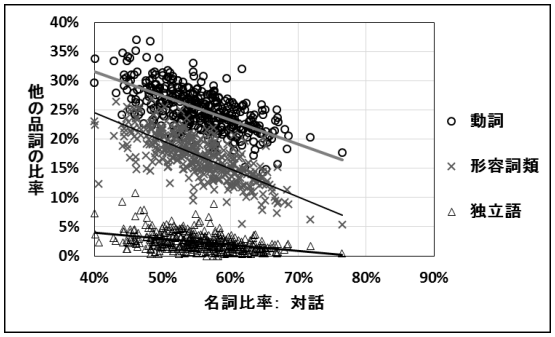


図 7.3 名詞と他の品詞比率の散布図・対話

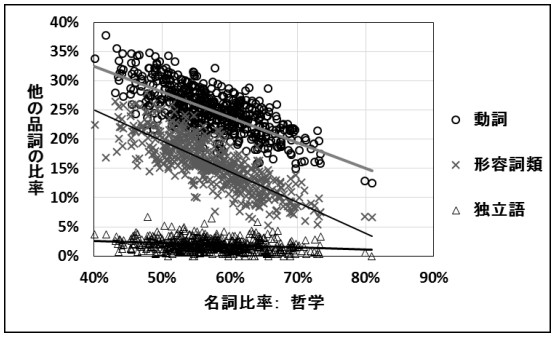


図 7.4 名詞と他の品詞比率の散布図・哲学

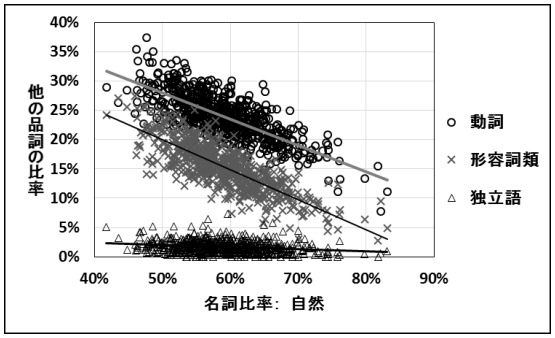


図 7.5 名詞と他の品詞比率の散布図・自然

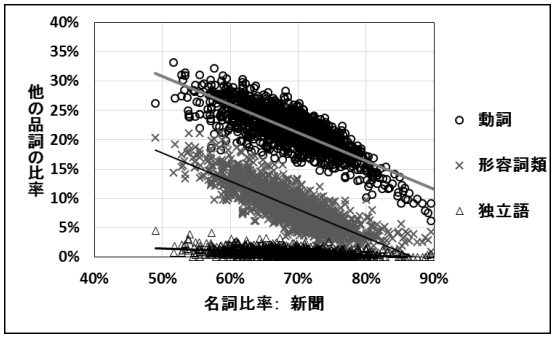


図 7.6 名詞と他の品詞比率の散布図・新聞

図 7.3～図 7.6 を見ると同じジャンルでも非常に大きなばらつき（分散）が存在し、一つの平均値で代表できるほどデータが固まって分布しているわけではないことが分かる。ただし、図 7.3 では名詞の比率が低い文書が多く、図 7.6 では名詞の比率が高い文書が多いなど、文書のカテゴリーによって名詞の分布に大まかな特徴が見られる。

表 7.1 図 7.3～図 7.6、図 7.8 の回帰直線の傾きと切片、並びに R^2 の値

	動詞			形容詞類			独立語類		
	傾き	切片	R^2	傾き	切片	R^2	傾き	切片	R^2
対話	−.413	.480	.414	−.481	.437	.539	−.106	.083	.167
哲学	−.437	.500	.533	−.527	.461	.646	−.036	.040	.050
自然	−.450	.505	.549	−.513	.456	.634	−.037	.039	.047
新聞	−.480	.548	.662	−.481	.419	.684	−.039	.034	.160
平均値	−.263	.398	.977	−.619	.515	.995	−.118	.088	.963

注：*t* 検定の結果、すべての傾きと切片は 5%水準で有意である。

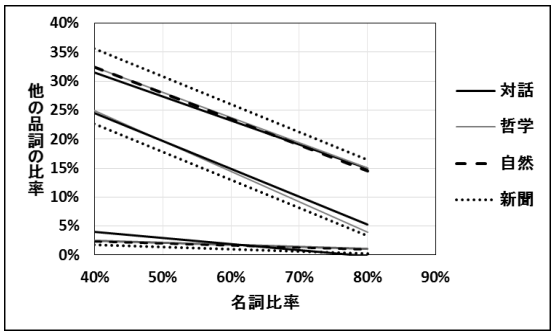


図 7.7 図 7.3～図 7.6 の回帰直線の比較

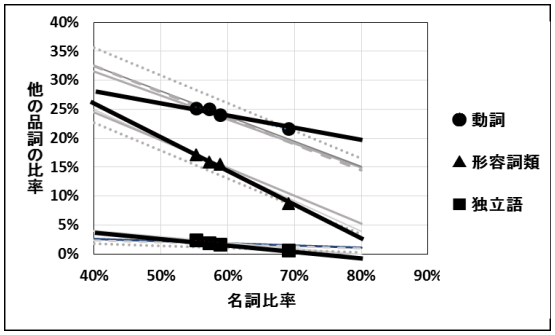


図 7.8 平均値を使用した回帰直線

図 7.7 は図 7.3～図 7.6 の回帰直線を比較した図である。動詞と形容詞類では図書館書籍の下位分類である「対話、哲学、自然」がほぼ同じで、「新聞」（点線）のみ傾向が異なっている。「新聞」は動詞が多く形容詞類が少ない。「新聞」はいつどこで何が起きたかという客観的な報道を中心に記述される媒体であるため、動詞が多く形容詞類が少なくなっていることが考えられる。独立語については、表 7.1 の R^2 値を見ても説明力が低く、どこまで有効な推定値になっているか疑問が残るが、「対話」のみやや傾きが大きくなっている。これは、「対話」の名詞頻度が低い文書において、感動詞が多く使用されているためだと思われる。

図 7.8 は各カテゴリーの平均値をマーカーで示し、回帰直線を描いた図である。マーカーは左から順に「対話」「哲学」「自然」「新聞」である。背景に薄く見えるのは図 7.7 と同じ回帰直線で、これと比較することにより、平均値を使用したことで、どれほど個体レベルの回帰直線からずれたかが分かる。図 7.7 では図書館書籍の三つのジャンルの回帰直線はほぼ同じ傾向を持っていた。つまり、個体レベルにおける名詞と他の品詞の相関の傾向性はほぼ同じであった。しかし、平均値を出すと名詞の頻度に若干の差があるため、図 7.8 のマーカーは横並びになっている。また、「新聞」の名詞比率の平均は、これらとは大きく異なるため、かなり離れた右側にマーカーが位置している。これらの

マーカーから計算された集団レベルの回帰直線は、媒体別の名詞率の平均値に強く影響されたグラフになっている。図 7.8 は、媒体の平均値という抽象化された値を使用して計算された回帰式であり、個体の観測値をもとに作られた回帰式ではない。このため、個体の観測値を当てはめてジャンルを判定しようとしても、判定精度が高くならないのである。図 7.8 を見ると、集団レベルの回帰分析と個体レベルの回帰分析では、異なる分析を行っていることがよく分かる。個体レベルの性質を明らかにする目的で、集団レベルの分析を行っても、有効な分析になりにくいことは明白である。

第 2 節　コーパスデータにおける生態学的誤謬と分割相関

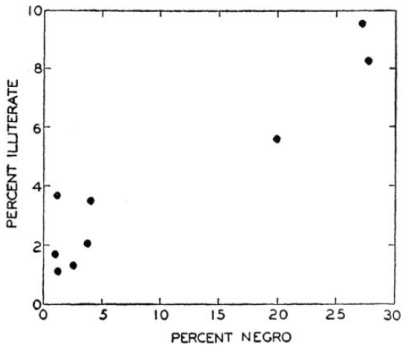
前節では集団レベルの回帰分析と個体レベルの回帰分析は、本質的に異なる分析であることを確認した。本節では集団レベルの分析では、本当は相関が低いのに相関が高いと誤認する生態学的誤謬や分割相関を見逃した分析によって誤謬が起きる例を示す。第 2.1 項では、生物学的誤謬と分割相関という用語の説明を行う。第 2.2 項では BCCWJ 固定長・長単位を使用して名詞と格助詞「の」の回帰分析を行い、これらの誤謬が起きる例を示す。

第 2.1 項　生態学的誤謬と分割相関の説明

生態学的誤謬は Robinson（1950）によってはじめて指摘され、社会学や政治学の分野で多くの議論を巻き起こした（森，1987:23 ; 北居，2014:133 ; 清水，2014: 1-2）。Robinson（1950）が取り上げたのは黒人の人口比率と文盲率などの相関で、これを地域別に集計したデータ（生態学的データ）で分析すると高い相関があるのに（図 7.9）、黒人や白人などの人間を観察単位にした個人データで分析すると低い相関しか見られない例などを挙げ（表 7.2）、回帰分析においてカテゴリーの合計値や平均値、比率などの生態学的データを使用して個体レベルの関係を推測する危険性を指摘した。

図 7.9 は、アメリカの九つの地域の黒人比率と文盲率を使用して描いた図で、ピアソンの積率相関係数は r = .946 と高い。表 7.2 は、図 7.9 と同様の分析を個体レベルで行った分析で、クロス表に使用されているのは人数である。このクロス表から計算された四分点相関係数 ϕ は、数学的には積率相関係数 r と同じ意味を持つ。個体レベルで分析すると、黒人と文盲率には ϕ = .203 という低い相関しか見られない。図 7.9 の集団レベルの分析結果から、「黒人は文盲率が高い」という個体レベルの推論を行うのは、誤りである。黒人が住んでいる地域の文盲率が高いからといって、黒人の文盲率が高いとは

限らない。このように地域で集約された割合のような生態学的データ（集団レベルのデータ）の相関に基づいて、個体レベルの推論を行った結果、誤った推測を犯してしまうことを生態学的誤謬という。



$r = .946$

Robinson（1950:338）表 1 より引用

図 7.9 人種と文盲の散布図

表 7.2 人種と文盲との個人相関

Robinson（1950:338）表 1 より引用

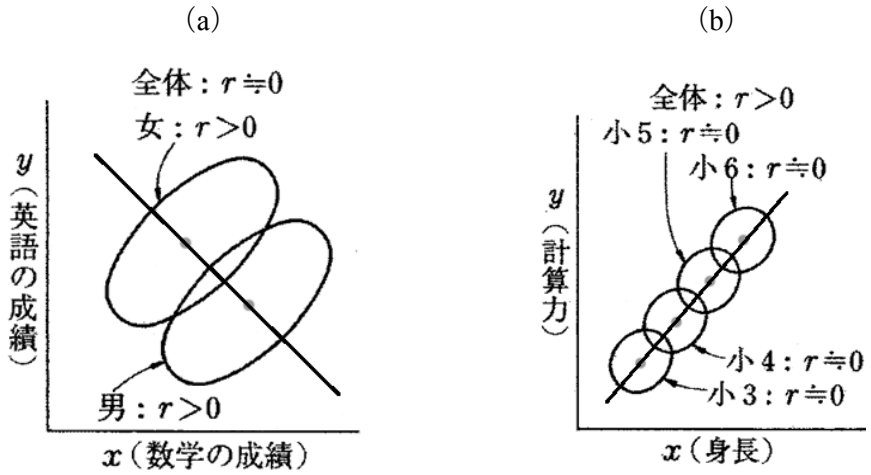
	Negro	White	Total
Illiterate	1,512	2,406	3,918
Literate	7,780	85,574	93,354
Total	9,292	87,980	97,272

$\phi = .203$

生態学的誤謬と同様に、集団レベルの分析と個体レベルの分析で異なった相関が現れる現象に分割相関がある。次の引用は、森・吉田（2013:228）による分割相関の説明である。

> 各ケースをいくつかの群に分けた時に，いずれか一方または両方の平均値や，両変数間の関係が，群間で異なっている場合（図 5 - 1 - 7）には，一般に群ごとに両者の関係を吟味する必要がある。このような群（層）別の相関を，分割相関（split correlation）または層別相関（stratified correlation）という。
>
> （森・吉田，2013:228）

図 7.10 の（a）は、数学の成績と英語の成績の相関散布図の概念図である。男女別で見ると、この二つの成績には相関があるが、男子生徒には数学の成績がよい生徒が多く、女子生徒には英語の成績がよい生徒が多いため、男女を合わせた全体では無相関になっている。男子と女子の平均値で回帰直線を描くと、負の相関が現れる。直線（負の相関）と全体（無相関）の関係が生態学的誤謬、全体（無相関）と男女別の楕円（正の相関）の関係が分割相関である。分割相関による誤謬は、個体レベルで相関関係を観察しているものの、その個体を適切な群に分割しないで全体で分析した結果、有効ではない分析結果となる誤謬である。



森・吉田（2013:228）図 5 – 1 - 7 より一部引用。平均値の印と直線は筆者が加えた。

図 7.10 分割相関の例

（b）は、身長と計算力の相関を小学 3 年生から 6 年生のデータを使って描いたグラフである。一般に、身長と計算力に相関があるとは考えにくいが、学年が上がるにつれて身長は高くなり、同時に計算力も上がっていくため、学年をまたいで分析すると相関が現れる。しかし、身長も計算力も一定の範囲に限定される同学年に分割して相関を求めると、それぞれの学年では無相関となる。（b）の場合は、学年の平均値で描いた回帰直線と、学年を混合した相関の傾向は一致しており、狭い意味での生態学的誤謬には当たらないが、分割相関を見過ごして集団レベルの分析を行った結果、身長と計算力に相関があると考えると不適切な推論となる。

第 2.2 項　生態学的誤謬と分割相関の例

集団レベルの分析だけを行った場合、生態学的誤謬を犯したり、分割相関に気づかないまま誤った推論を犯す可能性がある。これは、コーパスを使用した言語分析においても、十分に起こりうる危険性である。本項では BCCWJ 固定長・長単位を使用して名詞と格助詞「の」の回帰分析を行い、これらの誤謬が起きる例を示す。

丸山（2015）では名詞と「の」の関係について簡易的な分析が行われ、白書は名詞が多いことから「の」が多くなるという予想が述べられている（丸山，2015:139）。表 7.3 は丸山（2015）と同様の分析を固定長・長単位を使用して行った分析結果である。調整頻度は 1 文書当たりの語数に近い 500 語当たりの頻度で調整している。表 7.3 の名詞と「の」の調整頻度を見ると、最も名詞が少ない図書館書籍から白書にかけて徐々に頻度が増加し、これに連動して「の」の頻度も基本的に増加している。

表 7.3　BCCWJ 固定長・長単位における名詞と「の」の頻度

	図書館書籍	出版書籍	雑誌	新聞	白書
総語数	5,510,362	5,101,090	903,146	675,469	659,831
名詞の粗頻度	1,531,616	1,517,273	305,944	252,697	269,478
「の」の粗頻度	340,436	326,835	55,194	48,925	63,085
文書数	10,551	10,117	1,996	1,473	1,500
名詞の調整頻度:500語当たり	139.0	148.7	169.4	187.1	204.2
「の」の調整頻度:500語当たり	30.9	32.0	30.6	36.2	47.8

これを散布図で確認すると図7.11のように高い相関がある（r＝.844, df ＝3, p＝.072）。図中に示してある値はピアソンの積率相関係数 r である。図 7.11 の集団レベルの結果では丸山（2015）の記述は支持される。しかし、図 7.12 のように白書単体で個体レベルの散布図を描くと、名詞と「の」の相関係数は r＝.021（df ＝1,498, p＝.441）と無相関になる[26]。

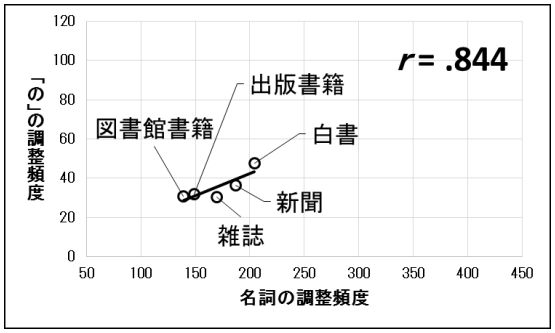


図 7.11 名詞と「の」: 媒体平均値

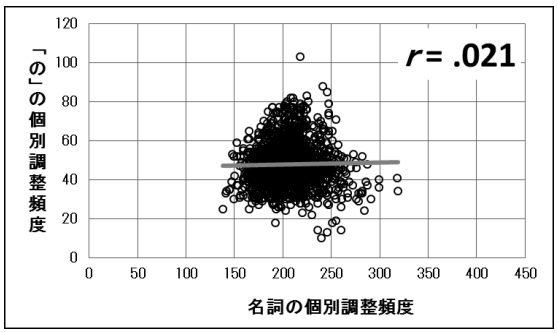


図 7.12 名詞と「の」: 白書

個体レベルで分析すると、白書は名詞が多いことから「の」が多くなるわけではない。図書館書籍や雑誌に比べて、白書で名詞や「の」が多いことは事実である。しかし、白書の文書を一つ一つ見ると、名詞が多く使われているからといって「の」が多く使われているとは限らない。図 7.12 の白書では、多種多様な文書が存在しており、名詞と「の」の間に何らかの関連性を見出すことは困難である。

「の」は、基本的に「体言＋の＋体言」の形で使用される。このため名詞に連動して「の」が増えるのはごく当然のように思われる。しかも図 7.11 のように媒体間で明瞭な傾向性を持っている場合は、個体レベルでも名詞と「の」の関連性を想定したくなる。

[26] 図 12 の r の p 値は，p = .441 で非有意である．相関係数の検定は r = 0 を帰無仮説としているため，これを棄却できないということは実質的に無相関であることを意味している。

しかし、このように当然と思われる現象でも、個別に分析してみるまで本当の状態は分からない。図 7.13〜図 7.16 は図書館書籍、出版書籍、雑誌、新聞を対象に、名詞の個別調整頻度と「の」の個別調整頻度で描いた散布図、図 7.17 は白書を含めた 5 媒体から各 1,400 文書を再サンプリングし、合計 7,000 文書で描いた散布図である。

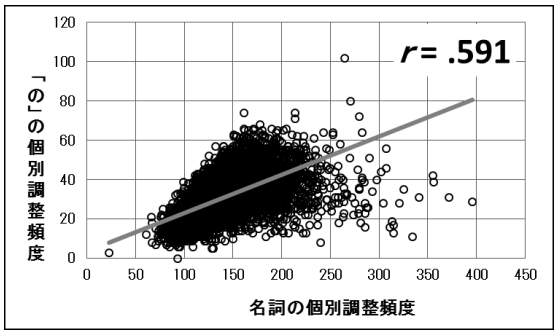


図 7.13 名詞と「の」: 図書館書籍

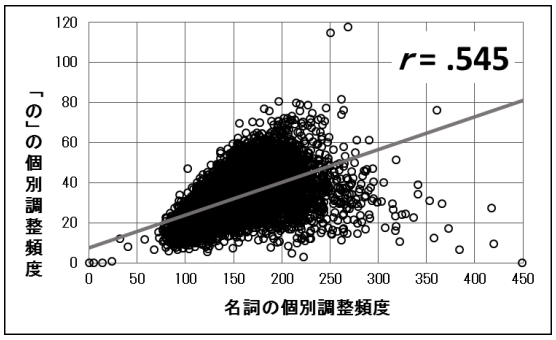


図 7.14 名詞と「の」: 出版書籍

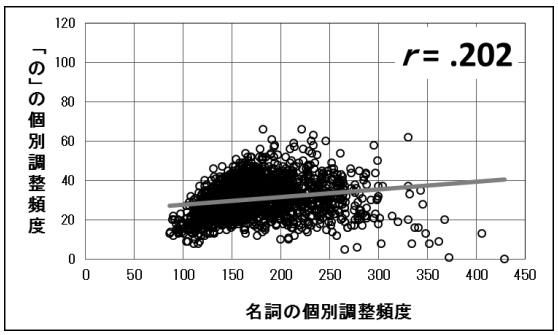


図 7.15 名詞と「の」: 雑誌

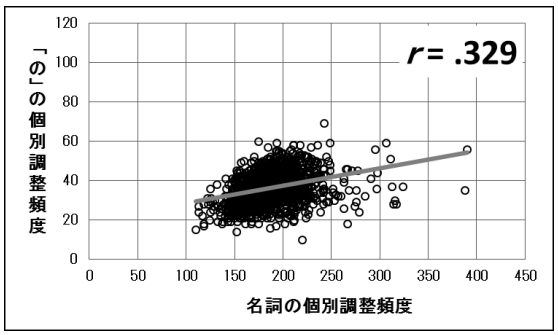


図 7.16 名詞と「の」: 新聞

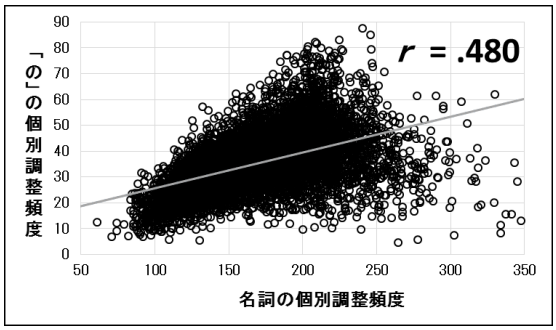


図 7.17 名詞と「の」: 5 媒体

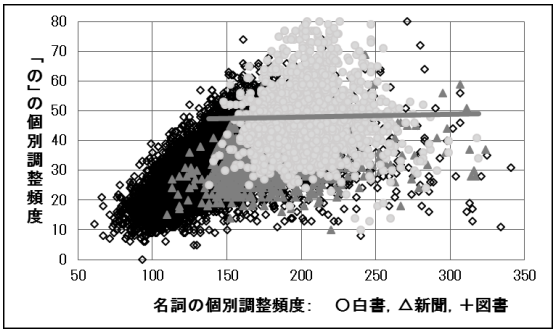


図 7.18 名詞と「の」: 3 媒体

これらの中で中程度の相関を示すのは図書館書籍（r＝.591, df ＝10,549, p＝0）、出版書籍（r＝.545, df ＝10,115, p＝0）、新聞（r＝.329, df ＝1,471, p＝0）の 3 媒体で、雑誌（r＝.202, df ＝1,994, p＝0）にはそれほど相関がない。

5 媒体を個体レベルで分析した図 7.17 の場合、集団レベルの相関係数の.844 から半分程度の.480 に下がる。この集団レベルの相関の強さから、個体レベルでも同程度の相関の強さを持っていると考えると生態学的誤謬を犯すことになる。

しかし、この例の場合、データの性質が異なる媒体を混合して図 7.17 のような分析を行うこと自体が、不適切な分析になっている。図 7.18 は分割相関が分かりやすいように、媒体数を三つに絞って描いた図である。図 7.17 のように媒体をまとめて全体で分析すると相関が見られるが、これは適切な分析ではない。図 7.18 のように群ごとに分けて描くと、群ごとに異なった性質を持っていることが分かる。

図 7.11 のように集団レベルで回帰分析を行った結果、白書の個体でも同程度の相関があると考えるのは生態学的誤謬、図 7.17 のように個体レベルの分析を行ってはいるが、異なる性質を持った群をまとめて分析した結果、適切な分析にならないのが分割相関を見逃したことによる誤謬である。生態学的誤謬を犯したり、分割相関を見逃して不適切な推論を犯すことを避けるためには、個体レベルで群ごとに分けた回帰分析を行う必要がある。

第 3 節　文書観察による変数の精緻化

本節では文書を観察単位として個体レベルの回帰分析を行うことのメリットを述べる。第 3.1 項では用例観察を行い、雑誌と白書ではなぜ名詞と格助詞「の」の相関が低いのかを考察する。第 3.2 項では、この考察に基づいて、変数を名詞から普通名詞に、格助詞「の」を普通名詞に接続する格助詞「の」に精緻化させることで、データのかく乱要因に左右されにくい分析が行えることを示す。

第 3.1 項　用例の観察

雑誌と白書ではなぜ名詞と「の」の相関が低いのだろうか。はじめに雑誌の文書を観察して、この理由を考察する。(5)は名詞頻度 1 位、(6)は名詞頻度 201 位、(7)は名詞頻度 501 位の文書の用例である。

(5) 『ＷＯＲＬＤ　ＭＥＧＡ‐ＢＡＴＴＬＥ　ＯＰＥＮ　ＴＯＵＲＮＡＭＥＮＴ　１９９９　ＫＩＮＧ　ｏｆ　ＫＩＮＧＳ　Ａブロック』１９９９年１０月２８日　東京・国立代々木競技場第２体育館▼トーナメント１回戦／５分×２Ｒ○
（PM41_00398，稲垣收・髙阪剛，『ゴング格闘技』）

(6) ブランソンには美しいゴルフ場が５つある。ここは千九百九十九年にオープンした最も新しいコース。　住Ｐ、Ｏ、Ｂｏｘ　二千二百七十三、Ｂｒａｎｓｏｎ、Ｍｉｓｓｏｕｒｉ　☎四百十七 - 三百三十九 - 四千六百五十三

（PM21_00680，中村嘉孝・今野憲雄，『ＤＩＭＥ』）

(7) 計６カ所の関節部分によりリアルなペダリングを再現。チェーンを模したゴムベルトで後輪を駆動しているあたりも自転車っぽいサイズ／長さ三百四十ｍｍ×幅百四十ｍｍ×高さ三百四十ｍｍ　カラー／シルバー、オレンジ

（PM21_00090，実著者不明，『ＣＹＣＬＥ　ＳＰＯＲＴＳ』）

これらを見ると、雑誌という媒体の特色がよく分かる。雑誌の文書数は 1,996 であるから、名詞頻度 501 位の(3)は名詞頻度上位からほぼ 1/4 の文書である。雑誌の中でそれほど特殊な文書ではない。それにも関わらず、一般的な文章という感覚とは大きく異なっている。これは雑誌という媒体が写真を多用し、記事の多くがカタログ的な文章となっているためである。これらの文書には文脈の中で名詞が現れるのではなく、情報提示のために列挙された固有名詞や数詞が数多く出現する。このため、名詞と「の」の相関が低いと考えられる。

次の(8)～(10)は、白書で名詞頻度 1 位、201 位、501 位の文書の用例である。

(8) 平成３年（年平均）の十五～二十四歳の青少年人口は千九百二十四万人で、このうち八百七十四万人（四十五。四％）が労働力人口である。　青少年の労働力人口を年齢階級別にみると、十五～十九歳が百八十三万人、二十～二十四歳が六百九十一万人で、二十～二十四歳が約８割を占めている。

（OW4X_00146，『青少年白書』，平成 4 年版）

(9) 勤労青少年スポーツ交流会は、昭和五十四年度から始められたもので、北海道、東北、北関東・新潟、南関東・甲信、北陸、東海、近畿、中国、四国及び九州の全十ブロックにおけるブロック内の勤労青少年ホーム利用者によるスポーツ交流会であり、種目はバレーボール、卓球、バドミントン、ソフトボール、テニス、スキー等のうちから１種目を選んで実施している。

（OW5X_00146，『青少年白書』，平成 8 年度版）

(10) 高等海難審判庁の裁決における事件種類別海難原因の主なものについて見ると，衝突事件に係るものが大部分を占めており、その中で海上衝突予防法の航法不遵守が最も多い（第八十三表）。第八十三表　事件種類別海難原因分類注　1　海難審判庁資料による。

（OW3X_00207,『交通安全白書』, 昭和 61 年版）

白書では (8)、(9)のように項目の列挙が多く、(10)の後半に見られるような表の表示や注が入る文書も多い。当然これらの名詞と「の」の頻度にも相関はない。雑誌や白書の用例を観察すると、これらの文書は図書館書籍に多い小説やエッセイなどの書籍や新聞の文書などとは大きく異なっていることが分かる。雑誌や白書には「の」が接続しない固有名詞や数詞が多く、図書館書籍に出現する名詞とは質的に異なっている。これが雑誌や白書で名詞と「の」の相関が低い理由だと思われる。名詞と「の」の関係を分析する枠組みにおいて、図書館書籍と雑誌や白書の名詞の頻度を同列に置いて比較することは困難である。

第 3.2 項　変数の精緻化

用例観察の結果、雑誌と白書で名詞と「の」の相関が低いのは、この 2 媒体に「の」とは無関係な固有名詞や数詞が大量に列挙されているためだと考えられた。そこで本項では変数を名詞全体から普通名詞とそれに接続する「の」に絞り込んで回帰分析を行ってみる。本研究ではこのように分析目的に対してより的確な変数に作り替えることを、変数の精緻化と呼ぶ。図 7.19～図 7.23 は、各媒体ごとに文書別の個別調整頻度を使用して描いた散布図である。

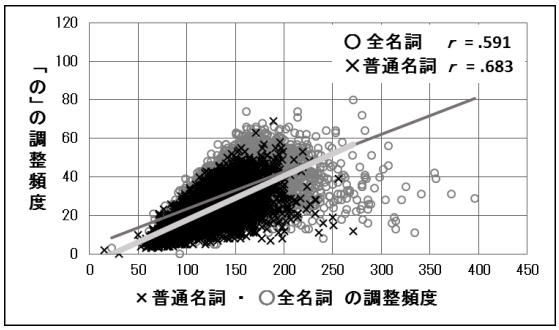


図 7.19 普通名詞と「の」: 図書館書籍

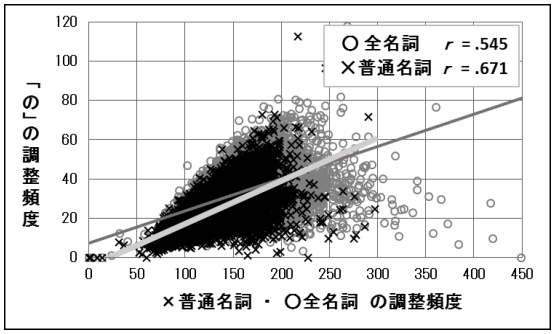


図 7.20 普通名詞と「の」: 出版書籍

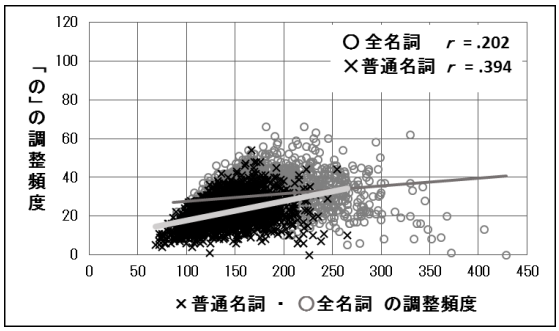


図 7.21 普通名詞と「の」: 雑誌

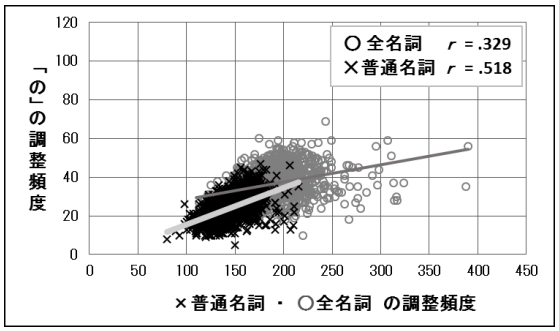


図 7.22 普通名詞と「の」: 新聞

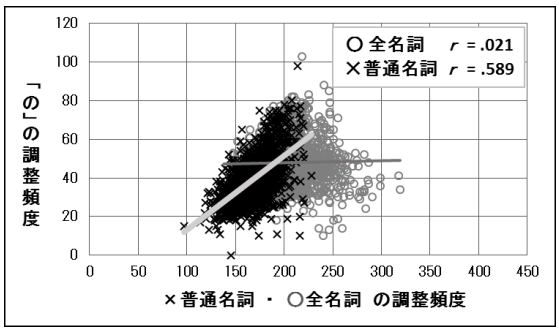


図 7.23 普通名詞と「の」: 白書

グレーの○印のマーカーが全名詞とすべての「の」の個別調整頻度で、図 7.12～図 7.16 と同じもの、黒い×印のマーカーがこれらから固有名詞や数詞を除いた普通名詞とそれに接続する「の」の個別調整頻度である。回帰直線は暗いグレーが全名詞、明るいグレーが普通名詞を使用して描いている。図 7.19 の図書館書籍では普通名詞に限定してもそれほど形や大きさは変わらない。ただし、ピアソンの積率相関係数は若干高くなる（r＝.683, df ＝10,549, p＝0）。図 7.20 の出版書籍も同様の傾向を示す（r＝.671, df ＝10,115, p＝0）。図 7.21 の雑誌では長方形に近かった形が図書館書籍に近い形になる。黒の普通名詞のマーカーに比べると、グレーの全名詞の部分が大きく、雑誌では図書館書籍に比べ多くの固有名詞と数詞が存在していることが分かる。普通名詞だけに限ると相関係数は r＝.394（df ＝1,994, p＝0）と高くなる。図 7.22 の新聞では、雑誌よりさらに普通名詞の面積が小さくなり、固有名詞や数詞が多く使用されていることが分かる（r＝.518, df ＝1,471, p＝0）。しかし、雑誌とは異なり、グレーの部分を含めた全名詞でも r＝.329 と一定の相関があるため、新聞の固有名詞や数詞は一定数「の」が接続して使われていることが分かる。新聞報道で出現する地名、人名などは、雑誌のようにカタログ的に羅列されるのではなく、文脈の中で使われるものが多いと考えられる。

図 7.23 の白書では全名詞と普通名詞で散布図の形が大きく変わる。全名詞では無相

関であったが、普通名詞に限ると r＝.589（df ＝1,498, p＝.441）と比較的高い相関が見られる。この結果から、図 7.12 の白書で無相関であった要因は、「の」が接続しない固有名詞や数詞が大量に存在しているためであることが確認できる。

あるコーパスでどのような現象が観察されるかは、そのコーパスにどのような文書がどのように分布しているかに起因している。このため統計的な言語分析では、回帰分析に限らず使用するコーパスに集積されている文書の観察が重要になる。文書を観察単位とし、個体レベルで行う分析法では、名詞や「の」の頻度といった言語情報を文書の単位で集約するため、特徴的な文書を探したり、頻度の高い順位の 1 位、200 位、500 位を抜き出して観察するなどのように、体系的に文書の観察を行うことが容易である。名詞と「の」の関係でいえば、名詞の頻度別に系統立てて文書を観察することにより、雑誌と白書では、分析のかく乱要因として大量の固有名詞と数詞が列挙されていることが観察できた。この結果、全名詞という変数を普通名詞に替えるという精緻化を行い、名詞と「の」との関係性の解明を前進させることができた。このように文書を観察単位とした個体レベルの分析を行うことで特徴的・体系的な文書の観察が容易になり、分析の精度を上げていくことができる。これが文書を観察単位として、個体レベルの回帰分析を行うことのメリットの一つである。

名詞と格助詞「の」の分析と、普通名詞と普通名詞に接続する「の」の分析では異なる分析を行っている。このため、普通名詞と普通名詞に接続する「の」の分析を行ったからといって、名詞と格助詞「の」の解明にはつながらないという批判が予想される。しかし、白書の全名詞のうち、82%は普通名詞である。また、全「の」のうち、93%が普通名詞に接続する「の」である。名詞と格助詞「の」の全データを使うことに拘泥して白書では名詞と格助詞「の」に相関はないと結論付けるより、普通名詞と普通名詞に接続する「の」に変数を精緻化させた方が、名詞と「の」の関係性の解明につながると考えられる。ただし、名詞と格助詞「の」の全貌を明らかにするためには、この分析に続いて、固有名詞や数詞とそれらに接続する「の」の分析を併せて行う必要がある。これは、今後の課題としたい。

第 4 節　分析対象となる文書の絞り込み　その 1

本節では、文書を観察単位とした個体レベルの回帰分析を行うことのメリットの 2 点目として、この分析法を使用すれば分析の目的に適した文書の絞り込みができるため、より研究目的に合った分析が行いやすいことを述べる。BCCWJ の固定長は現代語の使

用実態をありのままに反映させた均衡コーパスである。これは現実の言語の使用実態を解明する場合には貴重なデータであるといえる。しかし、その反面、文法や文体などの言語現象の解明に使用する場合は、そのままでは研究目的に適さない文書が存在し、分析が困難になることがある。本節では分析対象となる文書を絞り込んだ回帰分析を行うことによって、新たな知見が得られる可能性を示す。第 4.1 項では、本節の分析目的を述べる。第 4.2 項では分析に使用するデータの説明と文書を絞り込む基準を述べる。第 4.3 節では分析結果を、第 4.4 節では考察を述べる。

第 4.1 項　分析の目的

本章の第 1 節で取り上げた樺島（1955）は、日本語の品詞構成比率には一定の傾向性が存在することをはじめて明らかにした研究で、日本語の品詞グループの出現比率を定式化した数式は樺島の法則と呼ばれている。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 が增す」などの特徴が指摘されている。

これと類似の分析を古典語で行った研究に大野（1956）がある。大野が発見した古典語における品詞グループの比率は水谷（1965）によって定式化され、大野の法則と命名されたが、その後、樺島（2009：96）によって大野・水谷の法則と呼ぶことが提唱されている。水谷（1981）では大野（1956）の法則性を回帰分析として捉え直し、回帰式の精度を高める調整法が提案されている。大野（1956）の調査では、最も名詞比率の低い順から源氏物語、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、万葉集という順番で並び、その品詞比率も物語、日記、随筆など同じジャンルの作品が似た比率で固まることが明らかにされた。この特徴は樺島（1955：386）の主張と合致しており、日本語の文体的特徴と名詞には、関連性があることが解明された。

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・小西・小椋・小木曽ほか（2011）や山崎（2014）などがあるものの、名詞比率との相関は調査されていない。そこで本節では BCCWJ の固定長・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、名詞比率を説明変数、付属語を含めた品詞グループの比率を目的変数とした回帰分析を行って、名詞と付属語の関係を明らかにする。また、格助詞や接続助詞といった助詞の下位分類と名詞比率にどのような関

係があるのかについても調査する。

第 4.2 項　分析データと絞り込みの基準

分析には BCCWJ の図書館書籍と新聞、および CHJ の平安、鎌倉、室町編の長単位データを使用する。BCCWJ の固定長にはこの他に 3 種類があるが、白書や雑誌は前節で観察したように、大量の固有名詞や数詞の列挙が存在しているため、名詞と他の品詞比率を観察する本節の分析目的には適していない。また、出版書籍は第 6 章の格助詞「が」の分布から見ても、図書館書籍とほぼ同様の分布を示すことが考えられる。よって、使用データを図書館書籍と新聞に絞る。

BCCWJ と CHJ 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。本節では格助詞「の」を一時的に連体助詞という区分にして格助詞から分離して分類する[27]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[28]。

図書館書籍や新聞のデータを使用して、名詞比率を説明変数、助詞を目的変数とする回帰分析を行えば、名詞と助詞の関係が明らかになるはずである。しかし、実際にはうまく分析することができない。その理由は、BCCWJ の図書館書籍と新聞には多種多様な文書が存在し、中には分析目的に適していない文書も含まれているからである。図 7.24 の図書館書籍や図 7.25 の新聞の散布図を見ると、中央部の塊以外に、下部に散らばって存在する文書や尻尾のように分布する文書が存在することが分かる。これらは、第 2 章で観察した漢字カタカナ交じり文による解析ミスや図表の中身を言語データとして取り出した結果、名詞の列挙が多くなった文書など、文法や文体研究には適さない

[27] 本節では、格助詞「の」を他の格助詞グループから分離して分析するため、一時的に連体助詞と呼ぶ。

[28] 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。

文書である。図中の回帰直線の傾きがこれらの外れ値に影響を受けていることは明らかで、これらの文書を含めて分析を行う意義は乏しい。

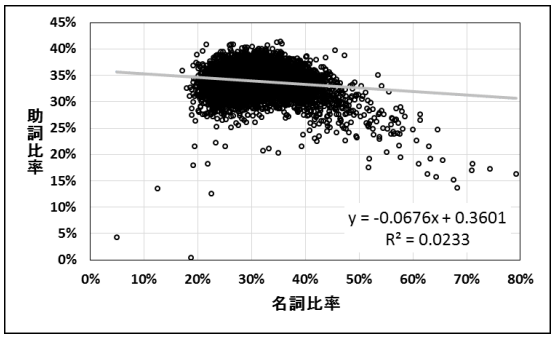


図 7.24 名詞比率と助詞比率の散布図

図書館書籍，N=10,551

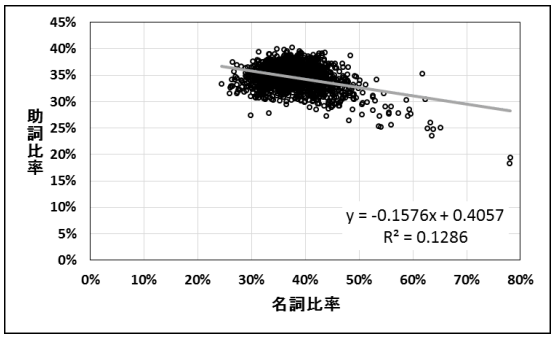


図 7.25 名詞比率と助詞比率の散布図

新聞，N＝1,473

その一方、CHJ では、古文の解析精度が高く、古典の作品が図表などを含まない文学作品にかぎられるためか、きれいな分布が観察できる。図 7.26 を見ると、*x* 軸の名詞の比率に対して *y* 軸の助詞比率は平行になっており、名詞比率が変動しても助詞比率が一定であることが分かる。一方、助動詞比率は、動詞やその他の品詞比率と同じような傾きを持つことから、文体的な特徴によって助動詞の比率が変動することが分かる。図 7.26 には外れ値となる作品が存在せず、CHJ には文法や文体の研究を行うのに不適切なデータは含まれていないと考えられる。

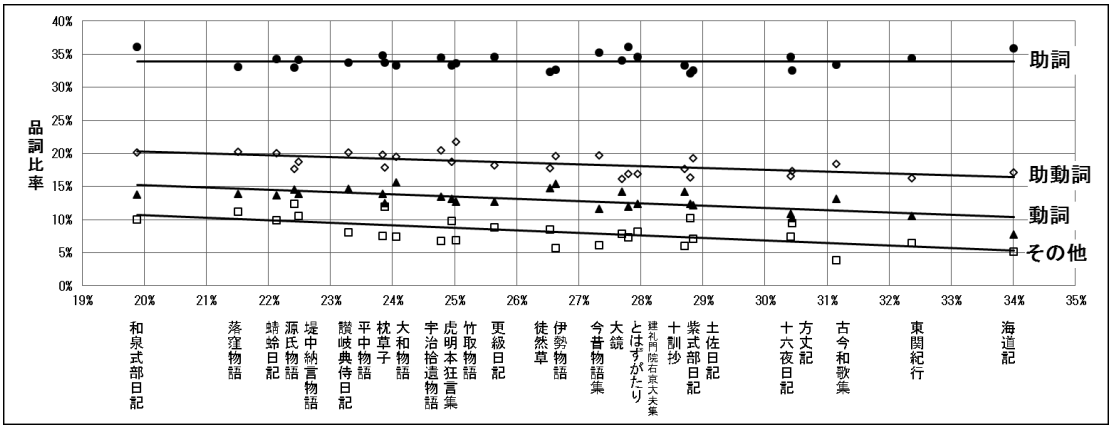


図 7.26 CHJ の回帰直線

図書館書籍や新聞でも、外れ値となる文書を除けば、研究目的に適する文書だけを使用した回帰分析が行えるはずである。しかし、1 万を超える文書から外れ値となる文書を個別に観察して判別していくことは難しい。形態素解析の誤りにしても、文書全体が

漢字カタカナ交じり文である場合は外れ値と認定しやすいが、文書のごく一部に漢字カタカナ交じり文が含まれている文書はどう判定すればよいのかなど、観察の労力だけでなく、外れ値を判別するための明確な基準を作ることも容易ではない。そこで本節では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、「漢字カタカナ交じり文」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「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。

第 4.3 項　分析結果

図 7.27 は図書館書籍に含まれる一般的な日本語テキスト 10,364 文書（残存率 98.2%）を使用して、名詞比率を説明変数、助詞比率を目的変数として回帰分析を行った結果を記した散布図である。図 7.28 は、同様の分析を新聞に含まれる一般的な日本語テキスト 1,353 文書（残存率 91.9%）を使用して分析した結果である。一般的な日本語テキストに絞り込んだ結果、古典語の分析と同様に助詞比率は一定であることが判明した。

図 7.29 と図 7.30 は、動詞、助動詞、その他の品詞の回帰直線を含めて、それらの回帰直線を描いた図である。これらを散布図で描くと、マーカーが重なって分かりにくいため、回帰直線だけで描いている。表 7.4 には、回帰直線の統計量を示した。

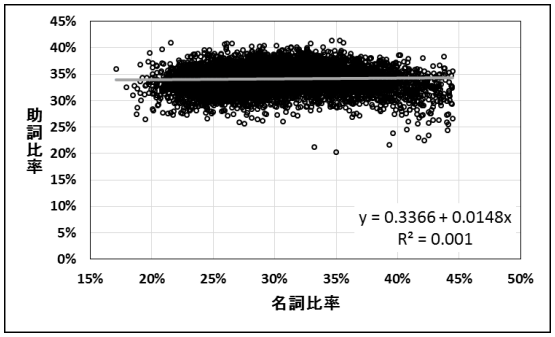


図 7.27 名詞比率と助詞比率の散布図
図書館書籍，N=10,364

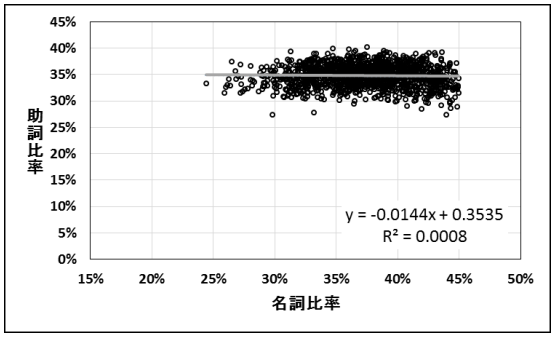


図 7.28 名詞比率と助詞比率の散布図
新聞，N＝1,353

これらを見ると、助動詞は動詞やその他の品詞グループ以上に傾きが大きく、名詞の増減に連動してその比率が大きく変化する品詞であることが観察できる。助詞と助動詞は名詞比率に対して、正反対の性質を持っている。

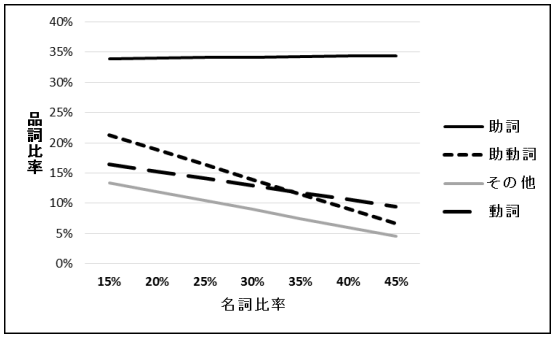


図 7.29 名詞比率と他品詞比率の回帰直線

図書館書籍，N=10,364

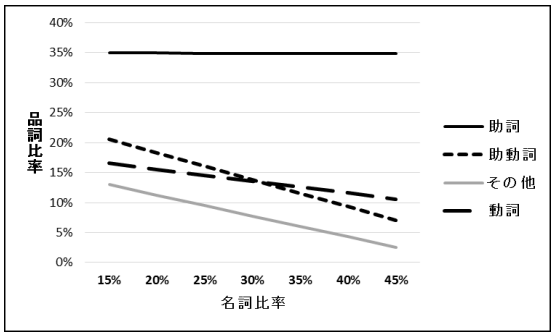


図 7.30 名詞比率と他品詞比率の回帰直線

新聞，N＝1,353

表 7.4　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

次に、名詞比率と助詞の下位分類の比率との分析結果を示す。表 7.5（図書館書籍）と表 7.6（新聞）は、自立語の品詞グループの比率と助詞の下位分類の比率を含めて調査した積率相関行列である。相関係数が±.500 超のマスに網掛けを施した。表中で結合助詞とあるのは、連体助詞と接続助詞を合計したグループを指す。これらは体言と体言、用言と用言を結合させる助詞であるため、仮に結合助詞と呼んでいる。これ以外の助詞は基本的に格関係の表示に関わる助詞と考えて格関係助詞という名称でグループ化している。

表 7.5　品詞頻度の積率相関行列：BCCWJ 図書館書籍の一般文書

	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
普通名詞	.814													
動詞	−.533	−.423												
その他	−.545	−.410	−.003											
助動詞	−.722	−.668	.222	.231										
助詞	.024	.089	−.106	−.326	−.381									
結合助詞	.340	.340	−.250	−.259	−.426	.429								
格関係助詞	−.261	−.195	.103	−.112	−.028	.652	−.405							
格助詞	.362	.415	−.007	−.480	−.419	.400	.036	.374						
係助詞	−.198	−.222	−.118	.087	.150	.223	−.190	.385	−.249					
終助詞	−.517	−.580	.230	.321	.365	−.003	−.325	.270	−.514	.005				
準体助詞	−.441	−.428	.133	.208	.286	.171	−.245	.379	−.238	.119	.425			
副助詞	−.193	−.045	.028	.157	.019	.184	−.159	.320	−.227	−.030	.159	.145		
接続助詞	−.549	−.486	.546	.199	.199	.156	.155	.027	−.285	.007	.349	.225	.098	
連体助詞	.650	.610	−.570	−.355	−.499	.274	.771	−.370	.215	−.170	−.508	−.359	−.202	−.509

表 7.6　品詞頻度の積率相関行列：BCCWJ 新聞の一般文書

	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
普通名詞	.631													
動詞	−.560	−.346												
その他	−.673	−.470	.110											
助動詞	−.714	−.664	.318	.456										
助詞	−.028	.278	−.098	−.304	−.486									
結合助詞	.050	.215	−.228	−.053	−.218	.373								
格関係助詞	−.063	.139	.055	−.277	−.351	.774	−.299							
格助詞	.228	.252	.097	−.534	−.425	.526	−.168	.655						
係助詞	−.442	−.407	.150	.432	.382	−.132	−.083	−.079	−.396					
終助詞	−.232	−.182	−.094	.230	.068	.204	−.160	.319	−.277	.056				
準体助詞	−.422	−.296	.127	.338	.264	.078	−.100	.148	−.199	.358	.175			
副助詞	.053	.304	−.073	−.118	−.237	.347	−.018	.369	−.028	−.019	−.122	−.026		
接続助詞	−.517	−.457	.340	.451	.386	−.142	.172	−.264	−.424	.403	.108	.232	−.105	
連体助詞	.363	.475	−.415	−.325	−.434	.423	.796	−.108	.109	−.323	−.210	−.233	.048	−.460

この表で注目されるのは、表 7.6 の新聞の一般文書において、名詞比率との相関が結合助詞比率：.050、格関係助詞比率：-.063 とほぼ 0 になることである。個別の分類では、接続助詞：-.517、連体助詞：.363、格助詞：.228、係助詞：-.442 と、名詞比率と一定の相関があるが、これらをグループ化すると相関がなくなる。この傾向は、図書館書籍では明確ではないため、何がこのような違いを生み出しているのか、この性質がどこまで一般化できるかを、今後詳しく解明していく必要があるが、助詞を計量的な性質によって、結合助詞と格関係助詞のグループに 2 分できる可能性が示唆されたことは、実に興味深い。

第 4.4 項　まとめと考察

前項の分析結果をまとめると、以下のようになる。

①名詞比率と助動詞比率には、強い相関がある。
②名詞比率と助詞比率には、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）[29]。
④格助詞や係助詞などと名詞には相関があるが、結合助詞以外の助詞の全てを「格関係

[29] 表 7.5 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」とはいえないが、表 7.6 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。

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）。

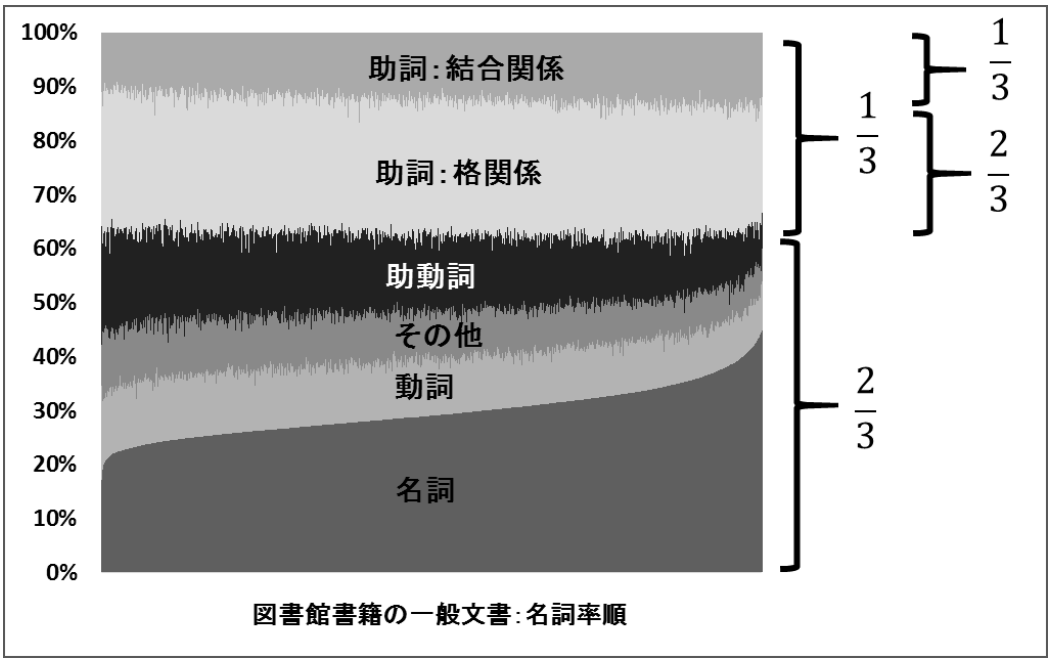


図 7.31　BCCWJ 図書館書籍に含まれる一般文書 10,364 の品詞比率の割合

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。これを分かりやすく図示すると、図 7.31 のようになる。この図は、図書館書籍に含まれる 10,364 の一般文書について、それぞれの品詞比率を算出し、名詞比率の低い順からソートして棒グラフを描いた図である。一見、面グラフのように見えるが、棒グラフが 10,364 本並んでいる図になっている。

これまで、名詞比率と自立語の動詞比率や形容詞類の比率などに相関関係があることは知られていたが、名詞比率と付属語の関係は不明なままだった。本節の調査によって、名詞比率と助動詞比率には他の自立語よりも強い相関があること、一方、助詞比率には相関がないことがはじめて明らかになった。

BCCWJ という均衡コーパスが出現したことで、このような相関関係が簡単に調査できるようになったと思われがちだが、BCCWJ には文法や文体の研究には不向きな文書も含まれているため、BCCWJ を無条件に使用した場合、必ずしも有効な言語現象の解明につながるとは限らない。本節では、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、「漢字カタカナ交じり文」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「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。このような限定を行うことによって、BCCWJ の全データを使用した場合には発見が難しかった現象の解明につながった。このように、文書を観察単位とした個体レベルの回帰分析を行うと、分析の目的に適した文書の絞り込みができるため、より研究目的に合った分析を行うことが可能である。

第 5 節　分析対象となる文書の絞り込み　その 2

本節では、前節に引き続き、文書の絞り込みを行うことで、より精度が高い回帰分析が行えることを示す。前節では「名詞比率 45%未満・その他比率 30%未満」のサンプルを「一般的な日本語テキスト」と定義し、この条件に合った文書に絞り込んで回帰分析を行うことで、新たな言語現象の解明につながった。しかし、この絞り込みの基準は、あくまで筆者が設定した恣意的な基準であり、より分析結果に一般性を持たせるためには、妥当性の高い客観的な基準に従って絞り込みを行うことが望ましい。

そこで本節では文書を絞り込む客観的な基準を模索するとともに、第 3 節で使用した変数の精緻化を併用して、言語現象の解明を試みる。第 5.1 項では、本節の分析目的を述べる。第 5.2 項では分析に使用するデータの説明と文書を絞り込む基準を述べる。第 5.3 節では絞り込みの基準が有効に働いているかどうかを検証する。第 5.4 節では分析結果と考察を述べる。

第 5.1 項　分析の目的

本節では、格助詞「の」[30]と文体の関係を調査し、日本語のテキストで助詞比率がほぼ一定であるという現象の意味を考察するとともに、妥当性の高い客観的な基準に従って文書を絞り込む方法を検討する。これまで、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」、「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」などの特徴が指摘され、名詞比率の増減は、文体の違いに大きく関わっていると考えられる。

その一方で、どのような文書においても助詞の比率がほぼ一定であるということは、

[30] 前節では格助詞「の」を結合助詞の下位分類とする都合上、一時的に連体助詞と呼んだが、ここからはこれまでの呼び方に戻す。

助詞の頻度は文体とは無関係に使用されていることを示唆している。つまり、日本語を日本語として成立させるためには、どの文体でどのような助詞を使うかは別として、その総量としては、必ず言語量全体の約 1/3 を使う必要があると考えられる。ただし、一般的な日本語テキストにおいて助詞の比率はほぼ一定ではあったが、助詞の下位区分で見ると、格助詞「の」には名詞と正の相関があるなど、名詞比率に対して相関があるものがほとんどであった。それでは、下位区分の助詞と文体の関係はどうなっているのであろうか。たとえば、格助詞「の」の場合、硬い文体や客観的な文体では名詞の頻度が高いため、「の」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な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「の」の頻度が高くなるのであろうか。

「の」が文体とは無関係に使用されているとすれば、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、「の」は名詞に連動して増減するだけで、「の」の「総量」としての使用に人間の意志や個性はほとんど介在していない可能性が示唆される。もちろん、「の」の使用に当たって、人間の意志や個性が介在していないことは考えられない。たとえば「コーパスの分析の研究の発表」など、何度も「の」を繰り返すと意味が分かりにくくなる場合、「の」の繰り返しを避けて「コーパス分析の研究発表」と書き換えるなど、「の」の使用をコントロールすることは常時行われている。それにも関わらず、文書というテキストが一定量がまとまった単位で観察した場合、「の」の頻度が名詞に連動して出現するだけなのであれば、「の」の使用は、その文書を執筆した人間の意志や個性を超えた、量的な規則性に従っている可能性が考えられる。つまり、一定量の名詞が存在すると、それを意味のある情報として伝達するために、常に一定量の格助詞「の」を必要とする可能性である。これを明らかにするため、本節では格助詞「の」と文体の関係を調査する。

ただし、これを調査する場合、単純に名詞と「の」の頻度で回帰分析を行っても、言語学的に有効な結果が出ることは考えにくい。第 3 節では、雑誌や白書に固有名詞や数詞の列挙が多く、名詞と「の」にはほとんど相関がなかった。このため、変数を普通名詞と普通名詞に接続する「の」に精緻化させることで、これらの本当の関係性に接近することが可能になった。また、第 4 節の分析では、名詞と「の」の散布図において、多くの外れ値が存在していることが確認された。このため、これらを除いた一般的な日本語テキストで分析することによって、日本語の品詞比率の規則性を明らかにすることができた。本節の分析においても変数の精緻化や文書の絞り込みを行うことが必要だと考

えられる。ただし、その絞り込みの基準は、恣意的な基準より、より妥当性の高い客観的な基準であることが望ましい。そこで本研究のもう一つの目的を、この妥当性の高い絞り込み基準の検討に置く。

第 5.2 項　分析データと絞り込みの基準

分析には BCCWJ 図書館書籍の固定長・長単位を使用し、文体の認定には柏野（2013）の研究成果である国立国語研究所（2015）『BCCWJ 図書館サブコーパスの文体情報』（第 1 版）を利用する。ただし「専門度」だけは、日本図書コード（C コード）の第 1 桁の販売対象を使用する。C コードは書籍の流通のために作られた 4 桁のコードで、その第 1 桁は書店で本を陳列する際の目安を示す販売対象が記されている。これで「教養」「専門」に当たる文書は難易度や専門性が高く、「児童」に当たる文書は難易度や専門性が低いと見なすことにする。「専門度」の指標に国立国語研究所（2015）の分類を使用しないのは、C コードであれば、文体情報が付与されていない出版書籍や特定目的 SC の「ベストセラー」の文書でも使用でき、今後の分析を行う上で汎用性が高いためである。

次に絞り込みの基準について検討する。柏野（2013）では、図書館書籍の 10,551 文書に「専門度」「客観度」「硬度」などの文体的特徴を表す分類指標を付与するに当たり、文書を「文体判断が可能と判断されるもの」と「文体判断が単純にいかないと判断されるもの」の 2 種類に大別している。前者は「テキスト構造が単純なもの」（例：章節構造）、後者は「テキスト構造・紙面形式に特徴をもつもの」である。本節では仮に前者を普通文書、後者を特殊文書と呼ぶ。特殊文書に認定されている文書は、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、インタビュー等の「対話系」など、11 の観点から分類されている。特殊文書の文書数は、図書の 15.8%に当たる 1,664 文書であるという。次の(11)は「視覚表現多用系」、(12)は「データベースやリスト系」の文書の一部である。

(11) 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　（LBqn_00015，実著者不明，『昆虫』）

(12) 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Ｕｎｉ

ｔｅｄ　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　　（LBo5_00063，実著者不明，『世界遺産ガイド』）

(11)、(12)の文書では、「の」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写真のキャプションや項目のリストとして名詞の列挙が多く含まれるからである。これらの文書を除けば、より名詞と「の」の関係性を明らかにするのに適したデータに絞り込むことができる。このため、文書の絞り込みの基準には、国立国語研究所（2015）の分類を使用し、使用する文書を普通文書に限定することが考えられる。

しかし、この絞り込みを行っても、適切な文書だけに絞り込めるわけではない。

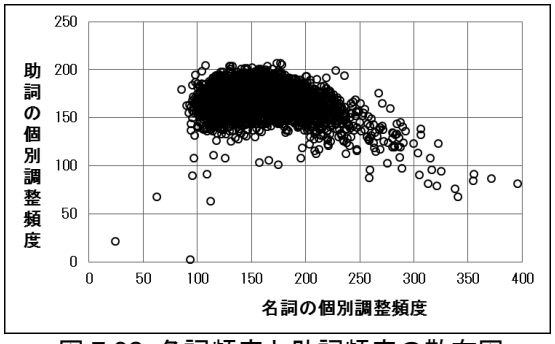


図 7.32 名詞頻度と助詞頻度の散布図

図書館書籍，N=10,551

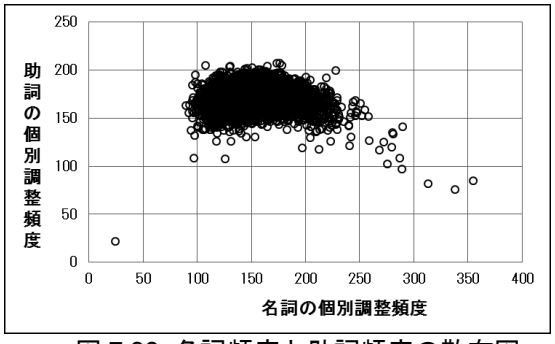


図 7.33 名詞頻度と助詞頻度の散布図

図書館書籍普通文書，N＝8,792

図 7.32 は図書館書籍の全データを使用して、名詞頻度と助詞頻度で描いた散布図、図 7.33 はこれを普通文書に絞り込んで描いた散布図である。柏野（2013:49）では特殊文書が 1,664 文書とされているが、国立国語研究所（2015）のエクセルデータで絞り込むと特殊文書が 1,759 文書となったため、ここでは普通文書の数を 8,792 としている。図 7.33 では、多くの外れ値が除去できているが、それでもいくつかの外れ値が散在し、分析目的に適した文書に完全には絞り込めていないことが分かる。どのような文書が残っているのか、用例を観察してみよう。（13）は図 7.33 で最も名詞の個別調整頻度が高いサンプル、（14）は散布図でマーカーの塊と尾の部分が分離し始める名詞の個別調整頻度が 250 語のサンプル、（15）は名詞の個別調整頻度が最も少ないサンプルである。

（13）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。　　（LBg6_00011，実著者不明，『首都圏白書』）

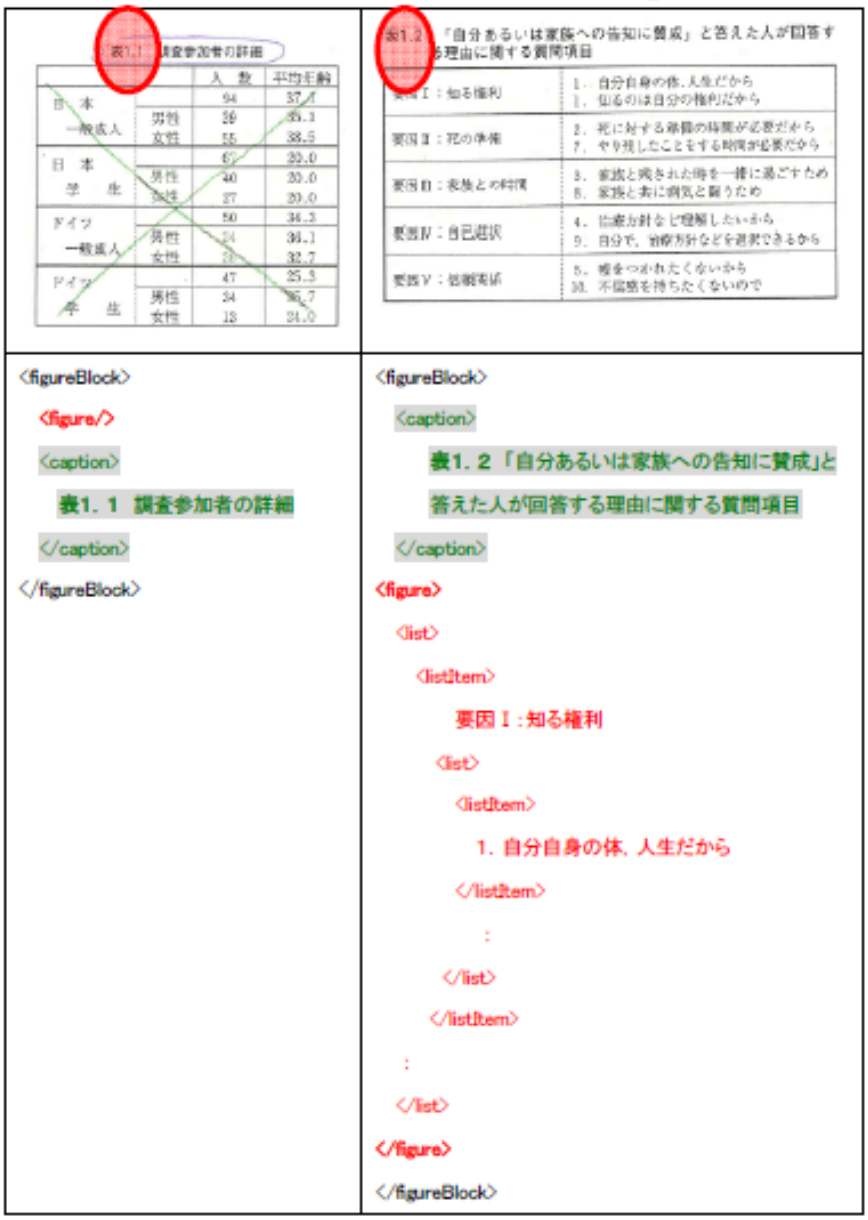
（14）あったら即食いの「ショコラドゥー」は、フランス・ヴァローナのチョコを使ったタルト。苦みの効いたブランドチョコを食べ続ける夢時間。三十メニュー公開食べ放題ＤＡＴＡ●料金／千二百八十円●制限時間／九十分●実施時間／十一：００～二十二：００（最終入店二十：三十）●種類／約三十種●飲み物／別・ドリンクバーはプラス二百円
（LBqn_00017，実著者不明，『Tokyo 噂の食べ放題』）

（15）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。
（LBs2_00005，松尾尊兌，『世界史としての関東大震災』）

（13）は、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。（14）は、雑誌で出現が多かった、カタログタイプの文書である。（15）は、漢字カタカナ交じり文で、形態素解析がうまくできていない文書である。これらの文書は、国立国語研究所（2015）の分類では、特殊文書には認定されていないが、名詞と「の」の関係性を分析する目的に合致する文書ではない。

そこで、分析目的に合致しない文書をさらに絞り込むため、BCCWJ の DVD 版に納められている M-XML ファイルに付与されている文書構造タグの利用を検討する。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 7.34 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。

〔PB24_00304：『生命倫理とこころのケア』〕



西部・大島・間淵・小林ほか，2011:271 より引用

図 7.34　図表からのサンプリング例

図 7.34 の左の表は、上段に表のキャプションがあり、その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キャプションのみである。

これに対し、右の表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。用例を観察した (14)は<figureBlock>が頻出する文書である。

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）の普通文書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。本研究ではこれを精選データと呼ぶ。

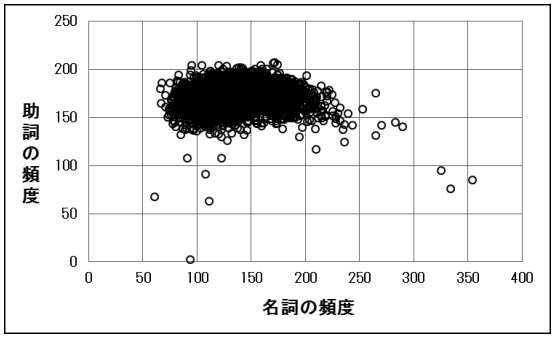


図 7.35 名詞頻度と助詞頻度の散布図：図書館書籍選抜データ：N=8,283

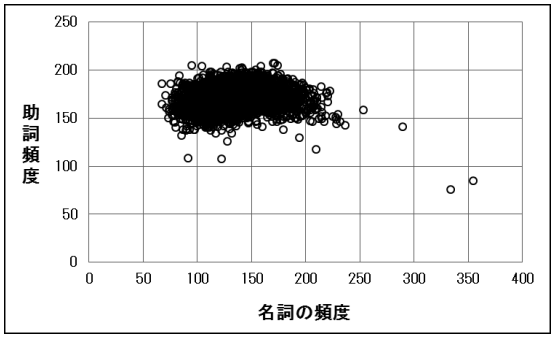


図 7.36 名詞頻度と助詞頻度の散布図：図書館書籍精選データ、：N=7,189

図 7.35 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 7.36 はここからさらに国立国語研究所（2015）の特殊文書を除いた精選データである。精選データになると、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、精選データでも、なお外れ値と思われるサンプルが若干残っている。

これを、今回の調査対象である名詞と「の」の散布図で描くと図 7.37 のようになる。この状態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 7.37 で名詞頻度が高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 7.37 で最も名詞が多いサンプルは、先に挙げた例文（13）の『首都圏白書』で、2 番目に多いものが（16）の『立川飛行場物語』である。これらは、写真のキャプションや図表の中身ではなく、文書の本文に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。

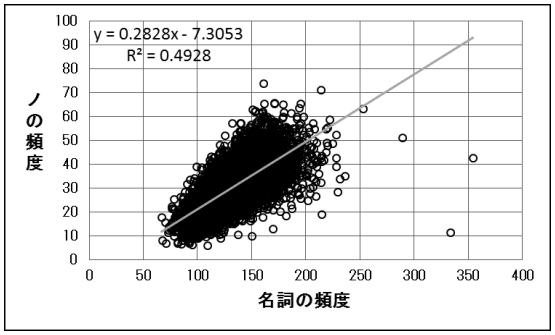


図 7.37　名詞頻度とノの頻度の散布図
図書館書籍精選データ：N=7,189

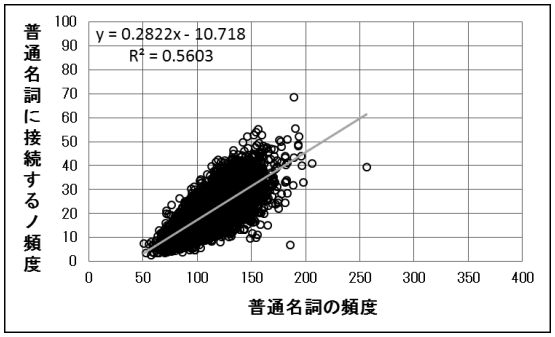


図 7.38　普通名詞頻度と普通名詞に接続するノの頻度の散布図
図書館書籍精選データ，：N=7,189

(16)　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　　（LBb3_00039，三田鶴吉，『立川飛行場物語』）

そこで図 7.38 のように名詞を普通名詞に、「の」を普通名詞に接続する「の」に精緻化すると、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精選データにし、普通名詞の個別調整頻度を説明変数、普通名詞に接続する「の」の個別調整頻度を目的変数とする回帰分析を行うことにする。

第 5.3 項　絞り込み基準の妥当性の検討

今回調査する文体は全部で 6 種類ある。この全ての文体別の回帰分析を行う前に、前項で検討した絞り込み基準が妥当かどうかを検討する。図 7.39 は、C コードの「教養・専門」と「児童」のフルデータで、普通名詞とそれに接続する「の」の散布図と回帰直線を描いた図、図 7.40 は、絞り込み後の精選データで同様の内容を描いた図である。フルデータを使用した図 7.39 では、「児童」より「専門・教養」の傾きが急で、専門度

の高い文書ほど「の」が多用されるという結果になる。しかし、精選データを使用した図 7.40 では回帰直線は一致し、「の」の使用に文体による違いはないという結果になる。

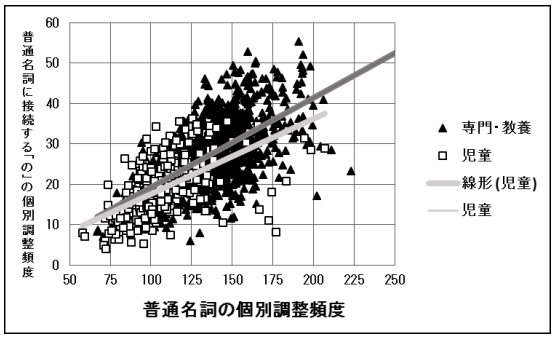


図 7.39 普通名詞と「の」: フルデータ

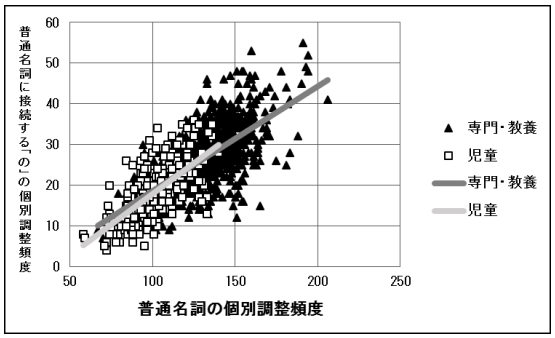


図 7.40 普通名詞と「の」: 精選データ

このどちらが正しい結果を表しているのかを検討するため、フルデータから精選データになるまで、一定数ずつ文書ずつ削除して描いた散布図を観察する。図 7.41 のフルデータでは、外れ値と考えられる文書が多く出現しているが、図 7.44 の精選データになるに従って外れ値が除かれ、分布の形状もスリムになっていくことが観察できる。

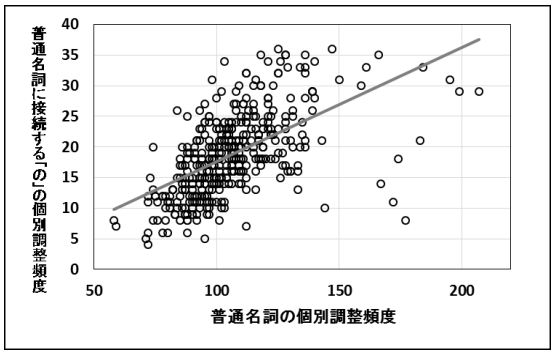


図 7.41 「児童」の散布図：フルデータ

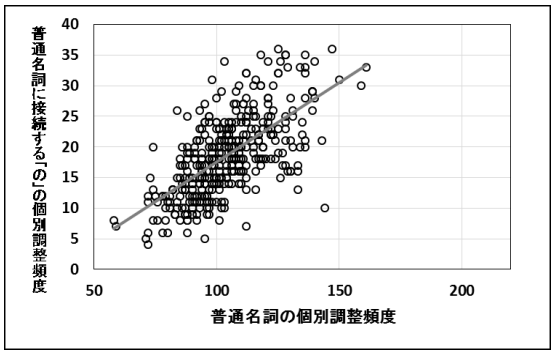


図 7.42 「児童」の散布図：10 文書削除

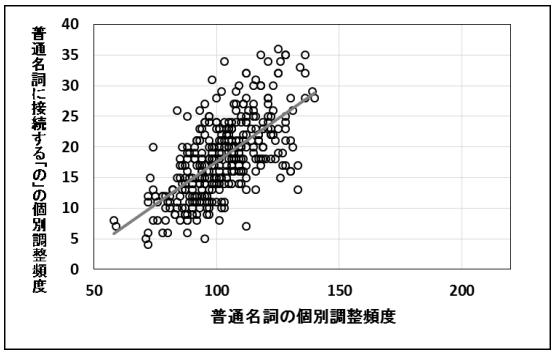


図 7.43 「児童」の散布図：30 文書削除

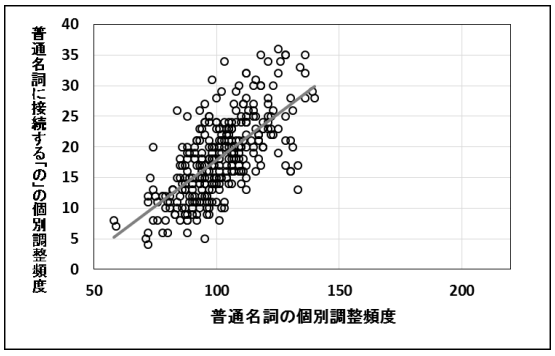


図 7.44 「児童」の散布図：精選データ

精選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(11)『昆虫』、(12)『世界遺産ガイド』が含まれており、(11)は「児童」、(12)は「教養」に分類されている。表 7.7 は、「児童」のフルデータ 387 文書から、分析に適さないと考えられる 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(11)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、章節構造を持った普通文書とは明らかに異なる構造を持った文書であることが分かる。

表 7.7：「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

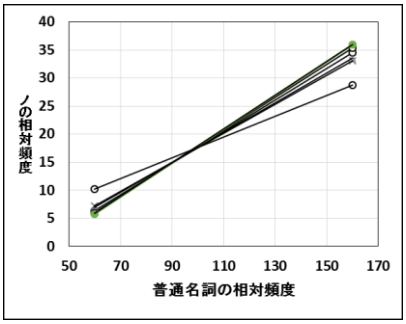


図 7.45 文書削除数別の回帰直線比較

ただし、<figureBlock>と<list>のタグがついている 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 7.45 は、フルデータから精選データまで、文書を 10 ずつ削除して求めた回帰直線を比較したグラフである。このグラフで回帰直線の傾きの変化が激しいのは、はじめの 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ができるかもしれない。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ことが妥当だと思われる。図 7.44 の精選データの散布図を観察すると、回帰分析に影響を与える外れ値は除かれており、概ねこの絞り込みの基準は妥当であったと考えられる。

以上の検討を基にすると、普通名詞とそれに接続する「の」の関係性には、専門・教養と児童という「専門度」の違いによって変化はないと考えられる。

第 5.5 項　五つの文体指標の分析結果と考察

前項の「専門度」の違いによる回帰分析に続き、「硬度」「くだけ度」「客観度」「語りかけ性度」及び話し言葉の「対話系」とそれ以外の文書における、普通名詞と普通名詞に接続する「の」の回帰分析の結果を示す。これらの指標は 2 段階～5 段階に区分され

ているが、図 7.46～図 7.50 では対極に位置する指標のみで分析している。表 7.8 は、これらの図に描かれている回帰直線の統計量である。

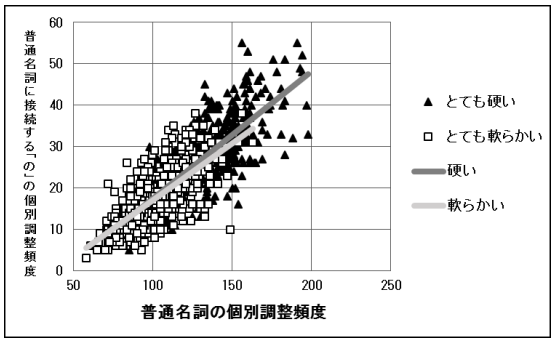


図 7.46 普通名詞と「の」: 硬度

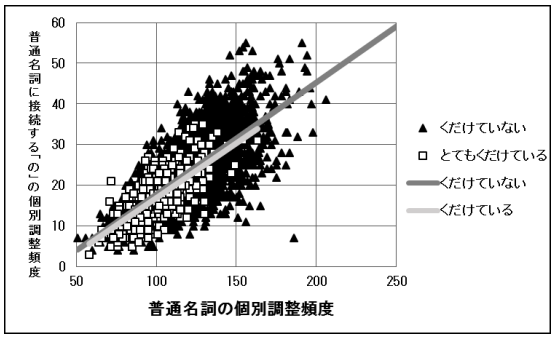


図 7.47 普通名詞と「の」: くだけ度

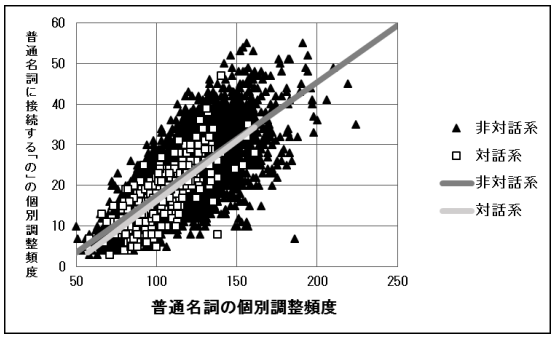


図 7.48 普通名詞と「の」: 対話・非対話[31]

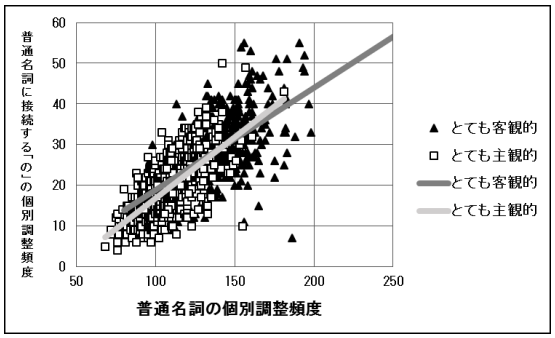


図 7.49 普通名詞と「の」: 客観度

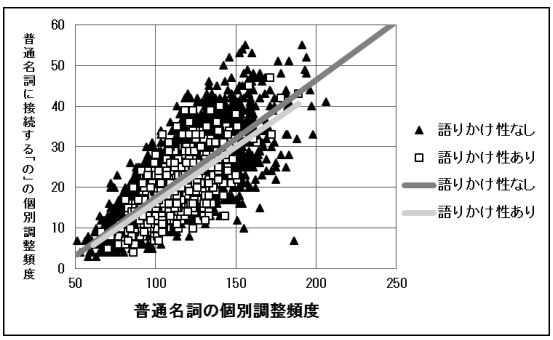


図 7.50 普通名詞と「の」: 語りかけ性度

表 7.8 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 7.46～図 7.50、及び表 7.8 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「の」の関連性は変化しない

[31] 「対話」は、特殊文書に分類されているため、図 7.46 は精選文書に「対話」を加えたデータで分析している。

ことが示唆された。

一方、「硬度」「くだけ度」「対話・非対話」では名詞の頻度によって明確な傾向性が観察された。文体が軟らかい文書、くだけている文書、話し言葉だけが使用されている文書は明らかに名詞数が少ない[32]。「客観度」と「語りかけ性度」でもこれらほど明確ではないが、主観的な文書や語りかけ性が強い文書で名詞が少ない傾向が見える。これらを観察すると樺島（1955:386）で主張されている「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」という指摘の正しさが追認される。樺島が指摘した特徴には、「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」などがあるが、図 7.46〜図 7.50 でもこれらの特徴が確認できる。

注目すべきはこれらの文体によって、普通名詞と普通名詞に接続する「の」の関係性にはほとんど変化がないことである。文体によって普通名詞の数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数が同じなら、使用される「の」の数はほぼ同じになる。つまり「の」によって連体修飾節を作る述べ方は、専門度だけでなく、話し言葉や書き言葉、文章の硬軟、くだけ度、客観度、語りかけ性度などには影響されないと考えられる。

この結果は、「の」の使用が、ある程度オートマチックに行われている可能性を示唆する。「の」を使用するに当たり、使用者が意識的なコントロールを行っていることに疑いはないが、一定量の普通名詞を使って情報を伝えようとすると、一定量の「の」が必要とされることもまた事実である。この問題は、日本語ではどうやって情報が伝達されているのかという、より大きな問題に関わっており、その解明が進むことで、日本語の一般的なテキストで助詞比率が一定である理由も明らかになると思われる。本節の分析結果を基に考えると、一定量の日本語で情報を伝達する場合、文体などの表現意図とは無関係に、全体の約 1/3 の助詞が必要であるため、一定になっていることが考えられる。この現象が、どのようなメカニズムで起きているのかの解明は、今後の課題としたい。

回帰分析は、外れ値に弱い分析法である。回帰分析を行うに当たっては、どうやって外れ値を除くかがカギになる。BCCWJ は、母集団の縮尺となるように多種多様な文書

[32] ただし，名詞数が少ない文書の文体が，すべて軟らかい，くだけている，話し言葉を使っているとは限らない．▲のマーカーは□のマーカーに隠れている部分が見えづらいが，名詞数が少ない位置にも分布している．

が集積されているため、現代日本語書き言葉の実態調査には向いているが、必ずしも文法や文体の研究に向いているわけではない。このため、BCCWJ を盲目的に使用するのではなく、研究目的に合った文書に絞り込むことが必要である。しかし、どのような基準で絞り込めば一般性を保った研究ができるかという絞り込みの基準は、簡単には決定しがたい。本節では、国立国語研究所（2015）という人手による文体の判断と、<figureBlock>と<list>という文書構造タグを使用して文書を絞り込む基準を検討した。<figureBlock>と<list>を使用したのは、一般的な日本語テキストとは認めにくい名詞の列挙を避けるためであり、分析目的が異なればこの基準が有効であるとは限らない。本節の検討ではこの基準がある程度有効に働いていると考えられ、この基準に従って文書を絞り込んだ結果、「の」の使用に文体の違いは関わっていないという新しい知見を得ることができた。しかし、この基準を使用したことで、必要以上に文書を削減してしまった可能性もあり、妥当な基準については、これからも引き続き検討していく必要があるだろう。

第 3 節の分析では、出版媒体によって「の」の回帰直線の傾きに違いがあった。これらの違いは本節で調査した文体指標以外の理由が原因だと考えるのが妥当である。これまで BCCWJ を使用した分析では、媒体別の比較を行って言語現象の特徴を論じることが多かった。たとえば、白書に特徴的な現象が観察された場合、ともすると白書の専門性やフォーマリティなどにその理由を求めることが多かった。しかし、少なくとも普通名詞と普通名詞に接続する「の」の関連性に関しては、それらの文体の違いが原因ではなく、白書では図表を多用して固有名詞や数詞が列挙されている文書が多いなどの言語学的な要因以外の理由が、回帰直線の傾きの違いを生んでいる可能性が高い。

合計頻度や平均値といった集団レベルで分析する場合、研究目的にふさわしくないデータを除くことは難しい。その結果、言語学的な要因以外で歪められている分析結果を、それとは気づかないまま受け入れてしまう可能性がある。一方、個体レベルの分析であれば、個別の文書を精査しながら、研究目的にふさわしいデータに絞り込んでいくことが可能である。このような文書の絞り込みが可能な点も、文書を観察単位とした個体レベルの分析法のメリットである。

第 6 節　まとめ

第 7 章ではコーパスで回帰分析を行う場合、文書を観察単位とし、個体レベルで分析を行うと有効な分析ができることを述べた。文字、単語、文などの言語単位を観察単位

とした分析法では、頻度情報に基づいた個体の観測値は全て 1 となるため、個体レベルの回帰分析を行うことは困難であった。このため、これまでの研究では、集団レベルの回帰分析が行われてきた。しかし、第 1 節で検討したように、集団レベルの回帰分析と個体レベルの回帰分析は、類似するように見えても、調査対象も分析結果も異なる分析である。また、第 2 節で観察したように、集団レベルの分析では、生態学的誤謬を犯したり、分割相関を見逃して誤った推論を行う可能性がある。

一方、文書を観察単位とすれば特徴的な文書や体系的な文書の観察が行いやすく、回帰分析の精度を高めていくためのメリットが生じる。第 3 節では、名詞数に着目して雑誌や白書の文書を観察した結果、これらの媒体では固有名詞や数詞の列挙が多いため名詞と「の」に相関がないことが判明した。ここから普通名詞とそれに接続する「の」で分析すれば、「の」の使われ方の解明がしやすいなどの考察が可能になり、変数の精緻化につながった。

メリットの 2 点目は、分析目的にそぐわない文書の削除が可能になる点にある。BCCWJ は均衡コーパスであるため、データが現実の縮尺になっているというプラスの側面と、そのままでは分析目的に不向きな文書も含んでいるというマイナスの側面を併せ持つ。回帰分析は外れ値に影響されやすい分析法であるため、分析目的に適した文書に絞り込んで分析を行うことが重要である。第 4 節では、日本語の品詞比率の規則性を調査するにあたり、「名詞比率 45%未満・その他比率 30%未満」のサンプルを「一般的な日本語テキスト」と定義して分析を行った。この結果、先行研究では未調査のままであった、付属語の比率と名詞比率の相関を明らかにすることができた。同じ付属語でありながら、助動詞比率は動詞比率のなどの自立語の比率よりも、名詞比率と強い相関を持っていた。一方の助詞比率は、名詞比率との相関はなく、一般的な日本語テキストでは全言語量の約 1/3 の助詞が使用されていることが分かった。また助詞の下位区分で名詞率との相関を調査すると、連体助詞（格助詞「の」）と接続助詞という語と語の結合を担う結合助詞のグループで助詞の 1/3 を占め、その比率もほぼ一定であることが判明した。

ただし、第 4 節で使用した絞り込みの基準は恣意的なものであったため、第 5 節ではより妥当性の高い客観的な基準を模索しながら、「の」と文体の関係を解明するための回帰分析を行った。基準には、国立国語研究所（2015）に基づいた人手による文体評定と、図表や項目の列挙を含む文書が特定できる文書構造タグを使用した。この 2 つを使用して文書を絞り込み、普通名詞と普通名詞に接続する「の」に変数を精緻化させて回

帰分析を行ったところ、「の」の使われ方には、「専門度」「硬度」「くだけ度」「客観度」「語りかけ性度」などの文体や話し言葉や書き言葉の違いには影響されないことが判明した。

以上のように文書を観察単位にした個体レベルの分析を行うことで、回帰分析のマイナス面を避け、プラスの側面を生かした分析をすることができる。回帰分析においては、文書を観察単位にし、研究目的に適さない文書の削除や変数の精緻化を行って、可能なかぎり純粋な言語現象が観察できる状態を作った上で、個体レベルの分析を行う分析法が有効だと考えられる。

第 8 章　結論

本章では、この研究全体を通すと何を明らかにしたことになるのか、全体の結論として何がいえるのかについて述べる。本研究の目的は、コーパスを使用した計量的な言語分析において、これまで当然視されてきた基本概念や基本的な分析法を再考し、文字、単語、文などの言語単位を観察単位と考えてきたこれまでの分析法に替わって、統計学的にも言語学的にも有効な分析法を体系的に提案することにあった。本研究の問いは「コーパスを使用した計量的な言語分析において、どのような分析を行えば統計学的にも言語学的にも有効な分析ができるのか」ということであり、その答えを 3 点に要約して述べれば、「①個体（文書や学習者）を観察単位として分析する、②分布図を地図として分析する、③かく乱要因に留意して分析する」ということである。第 1 節では、これまでのコーパス分析の課題と本研究の位置づけについて簡単にまとめる。第 2 節～第 4 節では、本研究の結論として①個体（文書や学習者）を観察単位として分析する意義と方法（第 2 節）、②分布図を地図として分析する意義と方法（第 3 節）、③かく乱要因に留意して分析する意義と方法（第 4 節）についてまとめる。第 5 節では、本研究全体の意義と今後の課題を述べる。

第 1 節　これまでのコーパス分析の課題と本研究の位置づけ

第 1 章ではコーパスをブラックボックスに例えた二人の研究者の言葉を引用した。

> このようにして近い将来に日本語のコーパスが広く使われるようになることは極めて望ましいことである。それを十分に活用するためには，それが存在するだけでは不十分であり，利用者の側にその活用に必要な知識と技能を得ようとする主体的な努力が要求される。コーパスは手軽に情報を得ることができるブラックボックスではないのであり，その性質を十分に理解した上で扱わなければ意味のある結論には結びつかないからである。　（後藤，2007:53）

> コーパスを統計処理するときに，一番さけたいことは，コーパスの内容も知らず，プログラムの処理内容も知らないままで，それらしい統計データを出すことである．いわば，ブラックボックスのコーパスをブラックボックスのプログラムで処理するわけであるが，その場合，それを行っている人間はいったい何をしたことになるのか．　（伊藤，2005:96）

ここから読み取れるのは、何の知識や技能もなしにコーパスを使用した分析を行っても、コーパスはブラックボックスにしかならないということである。コーパスのブラックボックスたるゆえんは、使い方がよく分からなくても、「手軽な情報」や「それらしい結果」が得られるところにある。しかし、コーパスの中身やその調べ方がよく分からないまま使っても、得られた情報や結果が何を意味しているのか、実のところは不明である。それにも関わらず、コーパスから得られた情報や統計結果を「意味のある結論」と結びつけて論じることが、少なからず行われてきた。この二つの引用はそのような風潮に対する警鐘である。

それでは、コーパスの活用に必要な知識と技能とは何だろうか。コーパスの設計内容や統計プログラムの処理内容を理解することだろうか。確かにこれらも重要な知識や技能には違いない。しかし、設計内容は、コーパスをどのように製作したかという製作者側の情報であり、完成したコーパスが実際にどのような「性質」を持つに至ったかは、コーパスを使用してはじめて明らかになることが多い。つまり、コーパスを使用しながらコーパスの性質を観察する技能も必須の技能といえる。さらにいえば、設計内容や使用観察で明らかになった性質も、単なる知識で終わっては意味がない。その性質に対してどのような対処を行えば有効な言語分析ができるのかという対処法が明らかになってこそ、有益な知識となり得る。また、統計プログラムの処理内容の理解にしても、最も必要とされているのはプログラミングの技能ではなく、自分が明らかにしたい研究目的に適しているのはどのような統計分析なのか、自分が用意したデータは、その統計処理を行うのに適切なデータとなっているのかが判断できる統計学の知識とその分析を適切に行っていくための技能であろう。つまり、どうすればコーパスの性質を把握できるか、その性質に対処するには、どんなデータを準備し、どんな分析を行えば、統計学的にも言語学的にも有効な分析ができるのかという問いに対する知識と技能こそが、切実に必要とされているのである。

しかし、これまでのコーパス言語学では、計量的な分析を行う上で最も基礎となるこのような知識や技能を明確にするための議論が、あいまいにされたまま研究が行われてきた。本研究は、統計の原理に合致した分析法とはどのような分析法なのか、どのような分析を行えば、言語学的に有効な分析ができるのかについて検討し、統計学的にも言語学的にも有効と考えられる分析法を提案した。本研究は、コーパスを使用した計量的な言語研究分野における最も基礎的な分析方法の原理と方法を体系化した研究に位置づけられる。

第 2 節　文書や学習者を観察単位とする分析法の意義と方法

コーパスを使用して統計学的にも言語学的にも有効な分析を行うために最も重要なポイントは、コーパスの個体（文書や学習者）を観察単位として統計分析を行うことである。文書や学習者を観察単位として分析する意義は、それが統計の根本原理に合致している点にある。本研究が最も重視する統計の根本原理は、次の 3 点である。

①母集団から無作為抽出された母集団の構成要素が個体である。
②個体は独立していなければならない。
③統計分析の目的は、個体の観測値の分布からデータの特徴や性質をつかむことである。

コーパスの研究では、これまで文字・単語・文などの言語単位を使用して分析が行われてきた。これは、コーパスの個体を言語単位だと考えていることを意味している。しかし、第 2 章で検討したように、Brown コーパスでも BCCWJ でも、実質的に無作為抽出されているのは出版物（の一部の文書）である。考え方によっては、文書の構成要素である言語単位も、集落抽出法によって無作為抽出されていると見なすことが可能かも知れない。しかし、文書の中で、文字や単語は一定の法則性に従って使用されており、独立していない。また、文字や単語の出現数にかかわる観測値は、出現したという意味で考えればすべて 1 であり、個体レベルでは分布しない。つまり、たとえ言語単位が無作為抽出されていると考えたとしても、言語単位を使用して有効な統計分析を行うことは難しい。

文書とは、コーパスを構築する際に抽出単位となった一定量のテキストのことである。この文書は、その抽出元である書籍や新聞などの性質を受け継いでいる。文書の性質を最も直接的に決定づけているのは、その文書を書いた執筆者だが、その執筆者も、どのような媒体にどのような目的で何を執筆するのか、執筆する文字数はどれぐらい与えられているかなどのさまざまな条件で書き方が変わる。つまり文書とは、ある執筆者がある条件を与えられたときに使用する言葉遣いを性質に持つサンプルである。このサンプルが大量にあれば、何らかの性質を持った執筆者が、何らかの条件を与えられたときに使用する言葉遣いが偏りなく集積できる。これが代表性を持ったコーパスである。

この文書を抽出するに当たって、ある文書の選択に他の文書の選択は何も影響を与えていない。つまり、文書は独立している。また、文書では、その文書に使用された言葉遣いを反映して、たとえば名詞が多く使用されているとか、助動詞が少なく使用されて

いるなど、観測値がそれぞれに異なる。つまり、文書の観測値は分布するため、この分布を基に分析を行えば、母集団の性質が合理的に推定できる。文書は、個体としての条件を満たし、文書を観察単位とした分析を行えば、統計分析の原理に合致した分析ができる。

広義コーパスの一種である学習者コーパスでは、無作為抽出が行われていない。このため、ある学習者コーパスから得られた分析結果は、その学習者コーパスだけに当てはまる結果であり、一般化することは難しい。しかし、学習者コーパスをどのような考え方に従って分析していけばよいかは、基本的に上記の分析法と同じである。

最も重要な視点は、学習者コーパスの個体とは何かを考えることである。学習者コーパスを製作する場合は、まず、学習者を集め、その学習者に作文を書いてもらったり、インタビューに答えてもらったりして学習者の言語データを集めている。コーパスによっては、一人の学習者から複数の作文や発話データを集めることがある。これらの言語データは、電子的に処理できるように文字化されるため、これらを文書と呼んでもいいだろう。

それでは、学習者コーパスの個体は、この文書なのだろうか。学習者コーパスでは、まず、学習者が選抜されている。学習者に作文やインタビューを課すのは、学習者の言語的特徴を反映したデータが欲しいからである。また、一人の学習者が複数の作文を書いた場合、それらの作文は共通する特徴を持っていて独立していない。このため、学習者コーパスにおける個体は、文書ではなく学習者と考えるのが妥当である。

学習者単位で集約した文書から、調査対象の観測値を合計すると、その値は学習者の性質を反映してさまざまに分布する。この分布から有益な情報を取り出すのが、統計分析の目的である。

コーパスにおける個体が文書や学習者であると考えると、カイ二乗分析で何を検定しているのかも理解しやすくなる。第 6 章で検討した表 6.1 は、格助詞「が」の使用率を BCCWJ 出版書籍と白書で比較し、カイ二乗検定を行った際の分割表である。これまでは表 6.1 のように、単語などの言語単位を使用してカイ二乗検定が行われていた。しかし、先に述べたように単語は文書の中で文法や文脈に影響されて出現するため、独立していない。表 6.1 は、7,404,994 語を一つ一つ調べて、それが「が」か「が」以外の単語かを調べたことを意味しているが、単語はいわば文書単位で数珠つなぎになってまとまっているため、一つ一つの単語が独立して取り出せるわけではない。

表 6.1 「が」の比較：出版書籍と白書（再掲）

	「が」	「が」以外	総語数
出版書籍	152,010	6,211,425	6,363,435
	2.39%	97.61%	100.00%
白書	16,888	1,024,671	1,041,559
	1.62%	98.38%	100.00%
合計	168,898	7,236,096	7,404,994
	2.28%	97.72%	100.00%

χ^2 =2364.841　p =.000　Cramer's V =.018

表 6.3　出版書籍と白書における格助詞「が」の文書度数比較（再掲）

	0～5	6～11	12～17	18～23	24以上	合計
出版書籍	354	2,322	4,248	2,537	656	10,117
	3.50%	22.95%	41.99%	25.08%	6.48%	100.00%
白書	339	588	415	132	26	1,500
	22.60%	39.20%	27.67%	8.80%	1.73%	100.00%
合計	693	2,910	4,663	2,669	682	11,617
	5.97%	25.05%	40.14%	22.97%	5.87%	100.00%

χ^2 =1204.243　p =.000　Cramer's V =.322

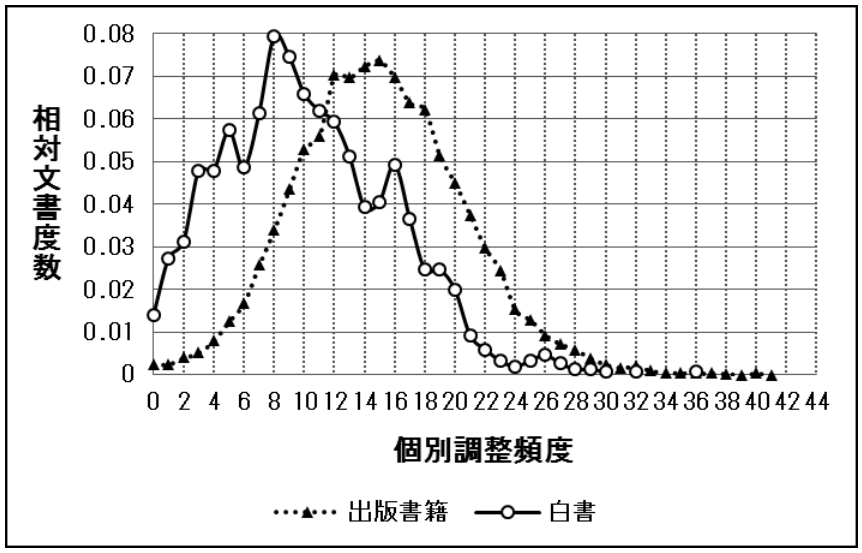


図 6.2 格助詞「が」の相対文書度数折れ線：出版書籍・白書（再掲）

表 6.3 の分割表における観察単位は文書である。文書に出現した「が」の頻度を 6 刻みで区分して、その区間に落ちる文書の度数を比較した表である。これは図 6.2 と同じように、媒体によって「が」をどれぐらい使う文書がどれぐらいあるかを比較している表である。図 6.2 は、表 6.3 の区間と図 6.2 の x 軸を対応させやすいように、右側に寄せて掲示している。このように、分割表と度数折れ線を対応させてみれば、表 6.3 を使用したカイ二乗検定が何を行っているのかは実に明瞭で、グラフの重なりの程度を検定

していることが理解できる。このような、個体を観察単位にしたカイ二乗検定であれば、クラメールの連関係数（Cramer's *V*）のような効果量も評価しやすい。

本研究の成果の一つは、以上のような、統計の根本原理に従って、何をコーパスの個体と考えればよいか、その個体の観測値を使用してどのような分析を行えば何が分かることになるのかという最も基本的な考え方の道筋を体系化して示したところにある。

今後、IT 技術の発展に伴って、新しいコーパスが次々と生み出されることが考えられる。しかし、それらを分析する場合にも、本研究で明らかにした考え方に基づいて判断すれば、統計学的な意義が明確な分析が可能になる。コーパスの分析法については、これまで当然のように使用されてきた分析法に対し、その分析法が統計の原理に照らして本当に適切なのかどうかを検討することが少なすぎた。本研究で示した、統計の原理に基づいた考え方を行えば、新しいテキストマイニングの技術などを取り入れていく際にも、その分析を行うことで、結局、何を明らかにしたことになるのかという最も重要な視点を見失わずにすむ。

第 3 節　分布図を地図として利用する分析法の意義と方法

コーパスを使用して統計学的にも言語学的にも有効な分析を行うために重要なポイントの 2 点目は、分布図を地図として利用する分析を行うことである。本研究で提案したコーパスの分析法は、分布図を描くことによって調査対象の全体的な特徴を把握しながら、最終的に対象に最接近して特徴的な文書の用例を観察する分析法である。これは Web で提供されている Google マップなどで、ごく粗い地図で全体のルートを俯瞰し、次第に目的の場所をズームアップさせて詳細に観察していく方法に似ている。

コーパスがブラックボックスになりやすい理由は、コーパスに集積されている大量の文書を、一つ一つ読んで特徴をつかむことが難しいためである。これは、数千万語もの言語データを読むことが労力的に大変なだけでなく、文書を読みながら全体的な特徴を把握することが困難であるためだ。一つ一つのテキスト（木）を見ても、その全体像（森）は見えない。コーパスを使用して調査対象の特徴を観察するには、まず、分布観察を行って全体像をつかみ、詳細に観察するべき文書を特定する。そして、どのような要因によって、その文書の頻度が決定されているのかという視点から、その文書のテキストを読んで特徴を把握する。何を読むべきか、どのような視点から読むべきかが明確になれば、より効果的に文書の細部を調査することができる。

この分析法については、第 4 章で詳細に論じたが、第 5 章の学習者の習得レベル別頻

度分布比較でも、第 6 章のカイ二乗検定を行う前の度数折れ線の比較でも、第 7 章の回帰分析を行う際の相関散布図の観察でも、全て同じ方法論によってコーパスの観察が行われている。以下、本研究で提示した代表的な分布図を見ながら、この観察法の意義を考えてみよう。図 4.1 は、BCCWJ 図書館書籍・固定長・長単位を使用して、格助詞「の」の度数折れ線を描いた分布図である。

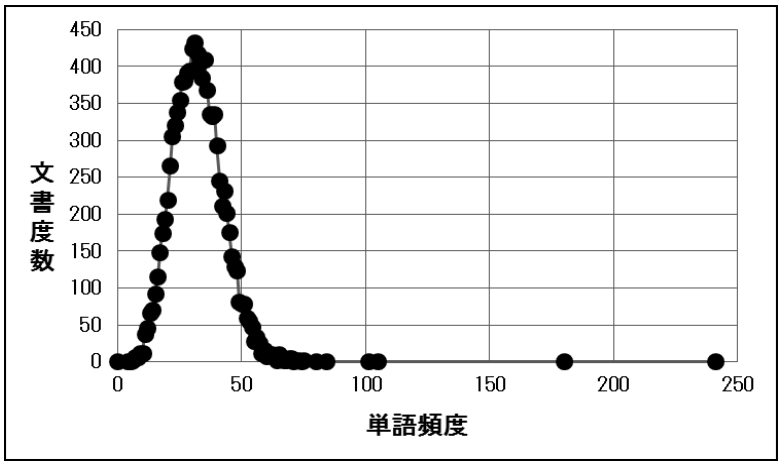


図 4.1　「の」の文書度数折れ線（再掲）

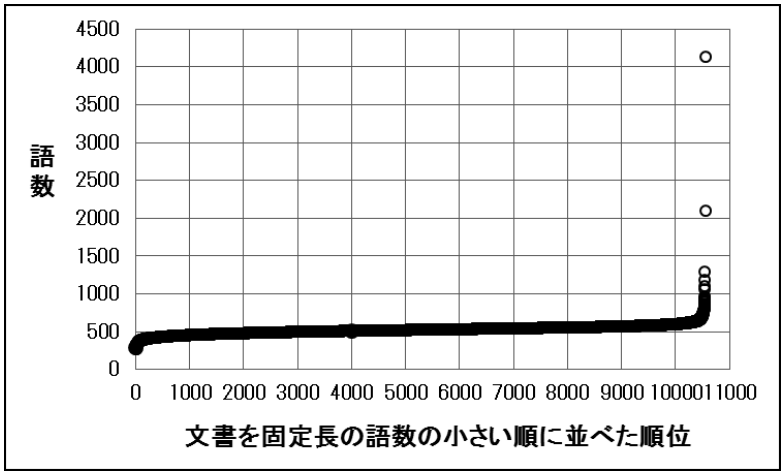


図 4.2 図書館 SC 文書の固定長語数（再掲）

固定長は、約 1,000 字に文字数を固定して集積された言語データで、統計的な分析に向くといわれている（丸山・柏野，2014:26；国立国語研究所コーパス開発センター，2015:30）。しかし、これは BCCWJ の設計段階で考えられていた内容であり、完成した BCCWJ をさまざまに使って検討した結果として述べられた内容ではない。図 4.1 を見ると、右側に外れ値が確認できる。これを使用して、統計分析を行ってよいのか、この

ようなデータは除いた方がよいのかについて、設計側からの説明はない。図 4.2 は、図書館書籍の 10,551 文書を語数の小さい順に *x* 軸に並べ、その語数を *y* 軸に描いた語数順と語数の散布図である。この図を見ると、なぜ図 4.1 の外れ値が出現しているのかが理解できる。固定長は約 1,000 字とはいうものの、中には異常に字数（語数）の多いデータが存在しているのである。そのような文書は、図 4.1 図だけ見ると、4 文書程度しかないように思われるが、図 4.2 を見ると、もっと多く含まれていることが分かる。

本来なら、このような情報は、BCCWJ のマニュアルや解説書に記述されていると、分析する側にとっては有益である。しかし、このような性質は、でき上がったコーパスがどのような状態になっているかという視点から観察して明らかになる場合が多い。分布図はコーパスを観察するための地図であり、地図を作ってはじめて、コーパスがどんな地形（性質）になっているのかという全体像が分かる。分布図の作成は、コーパス分析では欠かせない分析法である。

分布図を地図として利用する分析法では、文書の位置づけが把握しやすいことも大きな利点である。たとえば、用例を観察しただけでは、その用例が異常なのか、異常ではないのか、判断がつけられない場合がある。第 5 章の学習者コーパスを使用した分析では、KY コーパスの超級学習者で、条件表現の「たら」が多用されている用例を観察した。(2)がその用例の一部である。

(2) Ｔ：じゃあ、あのー、Ｓさんのね、あのふつうの日、授業がある日、〈はい〉の、いちにちーをちょっとー、朝起きてから、え、を一簡単に教えてもらえませんか

Ｓ：えっと一応ー、まあ授業がある一30 分前におきて、〈うんうん〉顔洗って、〈うん〉歯磨いて、〈うん〉服着てから、出て、いえを出て、〈ん〉んで学校に着いてからまあ、5 分余った**たら**、〈うん〉余ってる時間ジュースをちょっと飲んで、〈うん〉で、授業に入って、〈うん〉でまあそれでー1 講目 2 講目ー終わった**たら**、昼ー食べて、〈うん〉で 3 講目入って 4 こ目、終わった**たら**まあ 4 こ目終わってー、まあ、1 時間小 1 時間ぐらいー、えーっとリブレとか、〈うん〉あのー学食にいって、〈うん〉ちょっと食べって、〈うん〉友達としゃべって、〈うん〉んでーそれからー、まあ帰ってくるんですけれど [・・・・・]

Ｔ：うんー、あのーどんな番組を見てるんですか今

S：えーと一今はまあ、帰っ**たら**、結構夜遅くーなるんで、〈うん〉洋画とか、〈うん〉じゃなかっ**たら**、ああ、欠かさずに毎日見てんのは週間天気予報とか、　（KY コーパス、KS09）

この用例を読むと「たら」が多く出現するということは分かるが、だからといって特に使用方法に異常があるわけではない。しかし、分布図を描いてみると、他の学習者と比べてこの学習者がどれほど「たら」を多く使用しているかが明瞭になる。

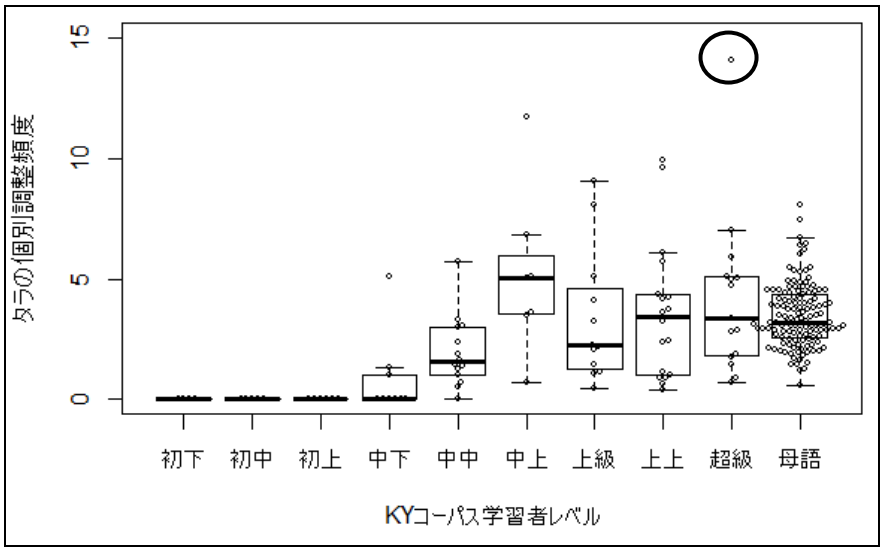


図 5.4　学習者レベルと「たら」の個別調整頻度の合成図。母語：名大コーパス（再掲）

図 5.4 は KY コーパスの習得レベルごとに、学習者ごとの「たら」の個別調整頻度を描いた蜂群図（マーカーが重ならないように描いた散布図）と、分布の傾向が分かりやすいように、箱ひげ図を重ね書きした合成図である。この図の黒丸で囲った部分が、(2)の学習者の個別調整頻度で、全学習者中、最も使用頻度が高く、突出している。この例を見ると、分布図はコーパスの地図であるという意味が理解しやすい。用例の観察は、現地に立って周りを観察しているのと同じである。しかし、それだけでは、自分が今、どこに立っているのかは分からない。分布図という地図と照らし合わせて、はじめて自分が立っている位置が把握できる。

これまでのコーパス研究では、質的な研究と量的な研究が別々に行われてきた。本研究が提案する分析法は、分布図という地図で全体を俯瞰するとともに、観察したい内容が的確に把握できる大きさまでズームアップし、最終的に用例という現地調査で質的な観察を行う分析法である。しかも、用例調査を行う場合は、あらかじめその用例の全体

的な位置づけが明らかになっているため、その用例のどこに注意して観察すればよいのかという分析視点が明確になっている。

次にズームアップの例を紹介する。図 5.8 は、I-JAS を使用して、図 5.4 と同じ方法で描いた合成図である。四角で囲った部分は、「たら」の使用が多い、個別調整頻度 5 以上の学習者である。

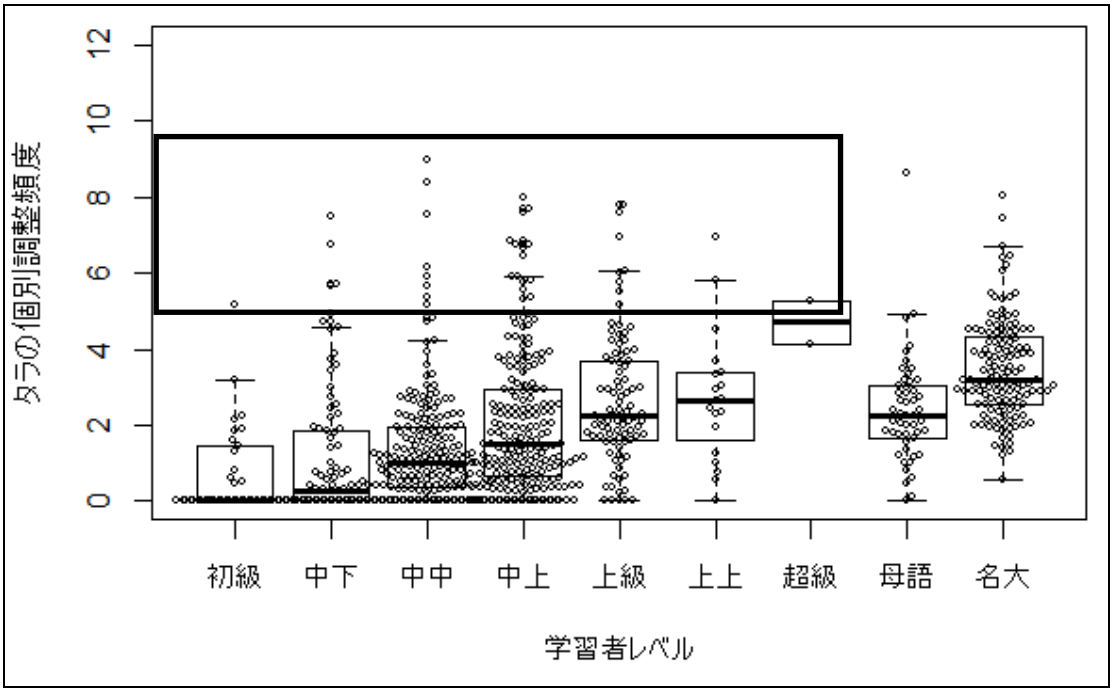


図 5.8 学習者レベルと「たら」の個別調整頻度の合成図、
母語：I-JAS、参考：名大会話コーパス（再掲）

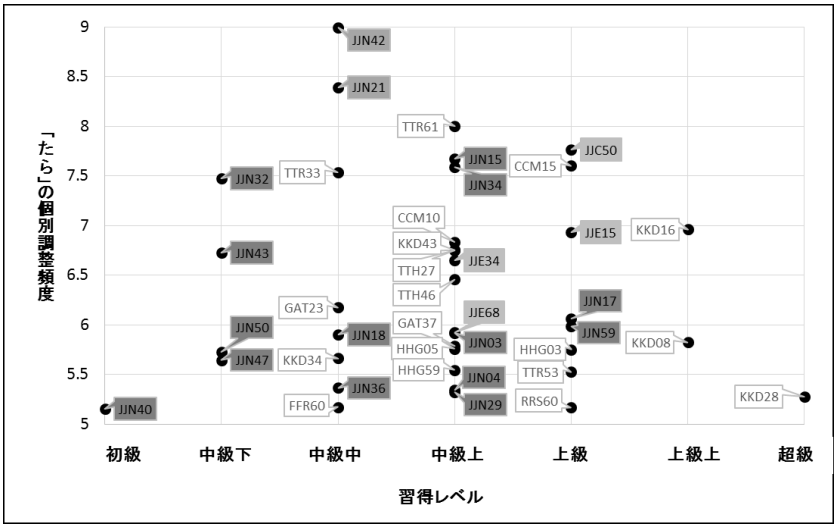


図 5.11 「たら」頻度 5 以上の学習者とその ID（再掲）

図 5.11 はこの部分をズームアップし、学習者の ID を併記した図である。国内の自然環境学習者（JJN）16 名は濃いグレーで、国内の教室環境学習者（JJC と JJE）4 名は明るいグレーで網掛けしている。I-JAS の第一次～第三次データの全学習者 610 名の中で、JSL 学習者はその 16.4%に当たる 100 名であるが、図 5.11 では、「たら」の個別調整頻度が 5 以上の学習者 39 名の半数に当たる 20 名が JSL 学習者で占められている。明らかに「たら」を多用している学習者には、JSL 学習者が多い。特に国内の自然環境学習者（JJN）は 50 名中 16 名が頻度 5 以上使用しており、習得レベルも初級から超級まで幅広いことが分かる。このことは、国内の自然環境学習者では、習得レベルが上がっても、「と」や「ば」の習得が進まず、「たら」に偏って使用する可能性を示唆しており、習得環境が学習者の言語習得にどのような影響を及ぼすのかを知る上でも、興味深い。

これまでの分析法では、習得レベル別の情報を、調整頻度で集約してしまうため、個別の学習者が見えてこなかった。本研究で提案するような分布図を作成すれば、問題のありかを切り取り、ズームアップして分かりやすく観察できる。ここからさらに個別の用例観察を行えば、より詳細な分析が可能になる。

以上のように、コーパスの個体である文書や学習者の分布図を描いて全体的な特徴を把握し、適宜、必要な倍率にまでズームアップして特徴を観察しながら、最終的に分布図で特徴的だった文書を特定してその用例を観察するという方法が、本研究で提案する分布図を地図として利用する分析法である。

第 4 節　かく乱要因に留意した分析法の意義と方法

コーパスを使用して統計学的にも言語学的にも有効な分析を行うために重要なポイントの 3 点目は、コーパスのかく乱要因に留意して分析を行うことである。本研究ではコーパス分析のかく乱要因として、外れ値、生態学的誤謬、分割相関を扱った。

このうち外れ値に留意した分析は、特に重要である。外れ値とは「データ集合の中で他の観測値と全く異なる観測値」のことである（Upton & Cook, 2011:301）。これまでのコーパス分析では、個体の分布図を作って観察するという方法自体がほとんど行われてこなかったため、外れ値の認定は用例観察によって調査対象外用例を特定するなど、質的に行われてきた。本研究の分析法では、観測値が分布するため、その分布の状態を見ることで外れ値が特定できる。外れ値は、合計や平均値、回帰直線などの分析に深刻な影響を与えるため、外れ値に留意しないと有効な分析はできない。

図 4.3 は、図書館書籍・短単位で調査した「色素」の文書度数折れ線である。「色素」

の合計頻度は 103 だが、15 文書中、一つの文書の頻度が 72 になっており、異常に高い。この用例を観察すると、色素名をまとめた表の中身が言語データとして抽出されたため、頻度が高くなっている。この文書は、異常値と考えられる。

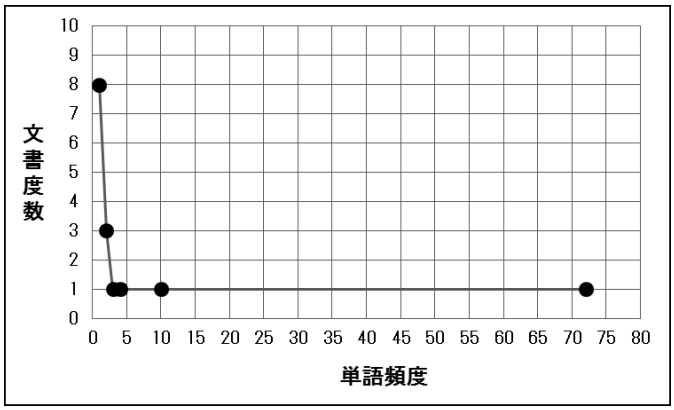


図 4.3　「色素」の文書度数折れ線（再掲）

BCCWJ の語彙表は、教育目的にも利用されているため、このような外れ値を含めた結果、「色素」という単語は頻度が高く、基本的な単語として教えるべきだと誤認される可能性ある。外れ値が全体の値に影響を与える場合は、これを除くか、文書の中身を再調査して頻度を決め直すなど、外れ値によって分析がかく乱されないように留意する必要がある。

学習者コーパスでは、このような外れ値が数多く存在し、合計値や平均値といった代表値に影響を与えることに注意が必要である。前節で取り上げた図 5.4 では、超級学習者の「たら」の個別調整頻度が、他の学習者から大きくかけ離れている様子を観察した。

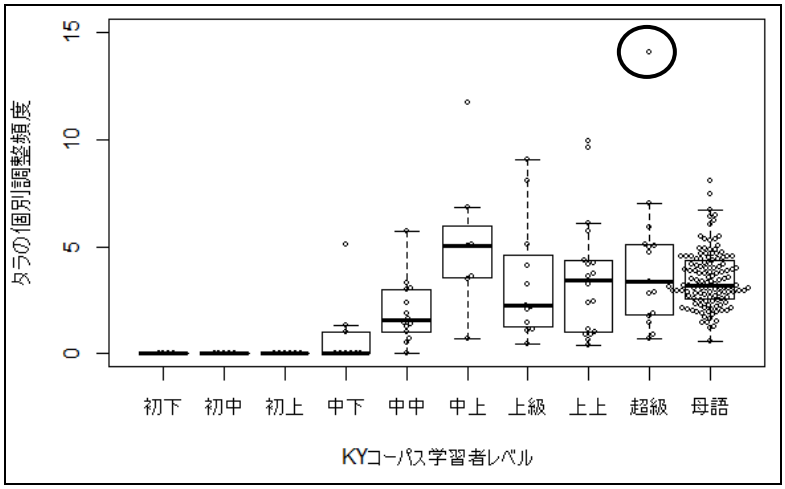


図 5.4　学習者レベルと「たら」の個別調整頻度の合成図。母語：名大会話コーパス（再掲）

この図で超級の 1,000 語あたりの調整頻度は 4.3、中央値は 3.4 である。どちらも代表値ではあるが、数値が大きく異なる。このどちらが正しいのだろうか。4.3 という調整頻度は、「超級の「たら」頻度合計÷超級の語数合計×1,000 語」で計算した値である。これは、一種の平均値であるから、この値が、外れ値によってかく乱されていることは明らかである。

このような外れ値は、超級以外のほとんどのレベルにも見られるため、他のレベルの調整頻度も適切な代表値にはなっていない。これまでの学習者コーパスを使用した研究では、基本的に調整頻度を使用して頻度比較が行われてきたため、これらの研究では正確な分析になっていない可能性が高い。本研究で提案したように、分布図を描いて分布を確認し、代表値であれば調整頻度ではなく中央値を使用することで、かく乱要因に留意した分析を行うことができる。

分布図を描く必要があるのは、中級上に見られるように、中央値が直ちに信頼できる値とはなっていない場合があるからである。中級上の場合は、このレベル全体に問題があると思われるが、このような判断は代表値として中央値を算出しただけでは判断が困難で、分布図を描くことで理解できる場合が多い。外れ値にかく乱されない方法は、分布図で分布を確認することと、外れ値が多ければ、合計値や平均値ではなく中央値を使用することである。

回帰分析も外れ値に影響を受けやすい分析法である。図 7.41 は、BCCWJ 図書館書籍・固定長・長単位を使用して、日本図書コード（C コード）の第 1 桁の販売対象で「児童」に分類される文書を使用し、普通名詞と普通名詞に接続する「の」で回帰分析を行った結果である。一方、図 7.44 は、人手による文体評定の情報（国立国語研究所，2015）と BCCWJ の文書構造タグを利用して、文体分析にふさわしい文書に絞り込んだ上で、同様の回帰分析を行った結果である。

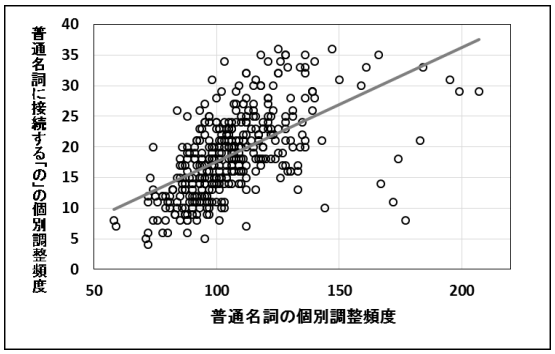


図 7.41 「児童」の散布図：フルデータ（再掲）

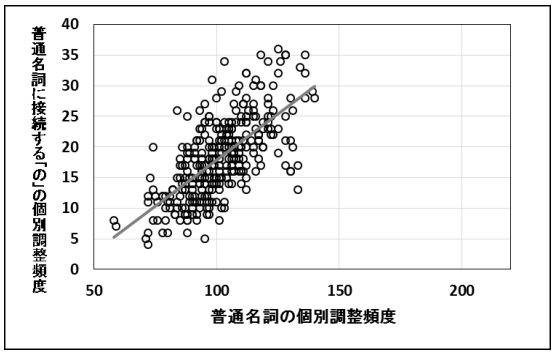


図 7.44 「児童」の散布図：精選データ（再掲）

図 7.41 を見ると、図の右側に外れ値が多く見られ、回帰直線がこれに影響されて傾きが低くなっているのが分かる。外れ値になっている文書は、図鑑やスポーツの解説書など、通常の文章とは異なって名詞が列挙されており、文体分析に適した文書とは考えにくい。このような文書は外れ値と考えて除いた方が、研究目的に合った分析ができる。図 7.44 は、適切と考えられる文書に絞り込んだグラフで、外れ値がなくなっている。

次の図 7.39、図 7.40 は、上の「児童」のグラフに、「専門・教養」のグラフを重ね書きして比較した図である。

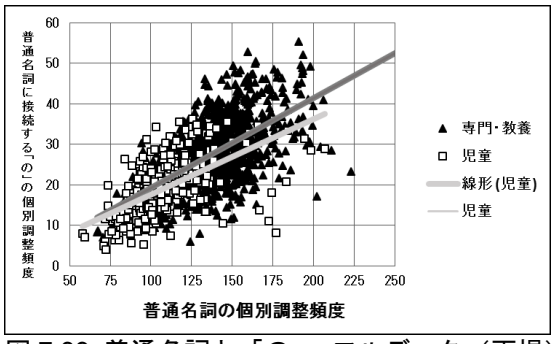


図 7.39 普通名詞と「の」: フルデータ（再掲）

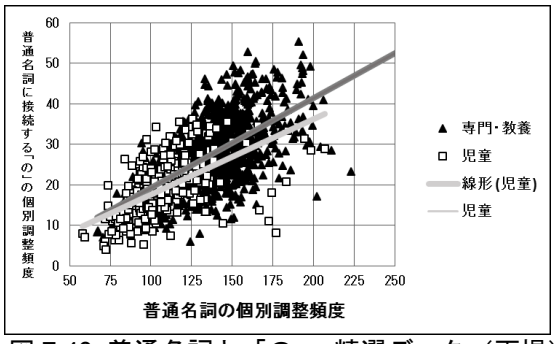


図 7.40 普通名詞と「の」: 精選データ（再掲）

外れ値の除去を行わない図 7.39 のフルデータでは、「専門・教養」という専門性が高い文書で「の」が多用されているという結果になる。しかし、外れ値を除いた図 7.40 では、「児童」か、「専門・教養」かという専門性の違いによって、「の」の使用に変化はないという結果になる。

注意を促したいことは、図 7.39 の結果が、いかにも「それらしい結果」に見えることである。専門性の高い文書ほど、複雑な内容を記していることから、連体修飾が多用され、「児童」に比べ、「の」が使われやすくなるという結果は、図 7.40 の「専門・教養」も「児童」も「の」の使い方は変わらないという結果より受け入れやすい。この例を見ると、これまでに行われてきた研究の結果がいかにも「それらしい結果」であったとしても、真実を示しているとは限らないことが分かる。

回帰分析では、生態学的誤謬や分割相関にも留意が必要である。合計値、平均値、中央値などの代表値は、集団の値を要約した代表値である。代表値を使用して学習者の習得レベル別の頻度を比較する分析は、集団レベルで頻度を比較することを意味している。回帰分析を行う場合は、このような集団レベルのデータの相関に基づいて、個体レベルの推論を行った結果、誤った推測を犯してしまう生態学的誤謬に注意が必要である。

図 7.9 は、名詞と「の」の調整頻度で描いた散布図で、ピアソンの積率相関係数は *r*

= .844 と高い値を示している。これは集団レベルの分析である。図 7.15 は、図 7.9 で使用した 5 つの媒体から 1,400 文書を再サンプリングし、合計 7,000 文書で描いた散布図で相関係数は r = .480 になる。図 7.9 の集団レベルの分析だけを行って、その結果から個体レベルでも強い相関があると考えることは、生態学的誤謬になる。

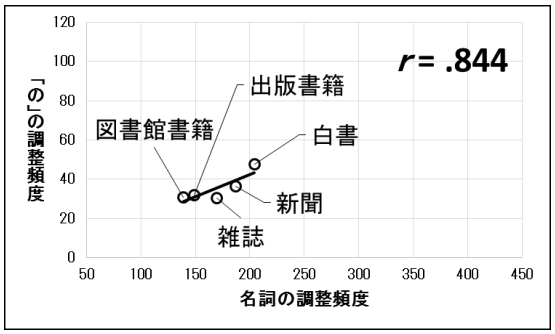


図 7.9 名詞と「の」: 媒体平均値（再掲）

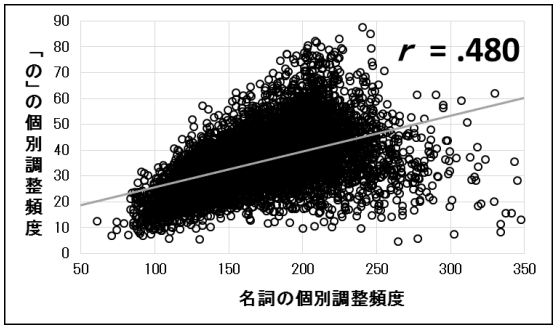


図 7.15 名詞と「の」: 5 媒体（再掲）

しかし、図 7.15 のような媒体を混合して分析すること自体が、実は不適切な分析となっている。それは、分割相関というかく乱要因を見逃して分析を行っているためである。図 7.16 が分割相関を分かりやすく示した図である。

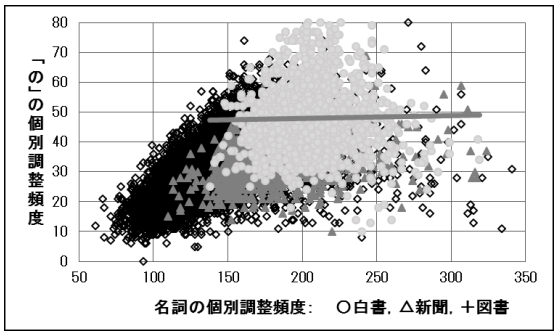


図 7.16 名詞と「の」: 3 媒体（再掲）

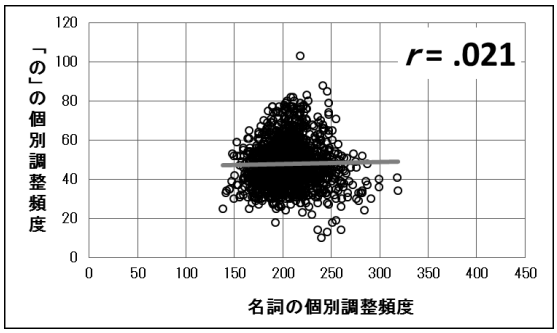


図 7.10 名詞と「の」: 白書（再掲）

図 7.15 の 5 媒体のうち、図が見やすいように図 7.16 では、白書、新聞、図書の 3 媒体だけで描いている。 図 7.15 のように 5 媒体を混合して分析すると、全体では r = .480 という相関を持っているように見える。しかし、図 7.10 のように白書という媒体を単独で調べると、r = .021 のように無相関である。図 7.15 は、図 7.16 で確認できるように、媒体ごとに異なった性質を持っている分割相関を見逃し、全体を混合して分析した結果である。この結果から、5 媒体の個体レベルで r = .480 という相関があると考えるのは誤った判断である。このように媒体ごとに異なった性質を持っている場合は、分割相関

というかく乱要因を見極め、それぞれ個別に相関係数を求めるのが有効な分析である。

コーパスを使用した分析では、分布観察を行い、外れ値、生態学的誤謬、分割相関などのかく乱要因に留意して分析を行うことが重要である。

第 5 節　本研究の全体的意義と今後の課題

本研究全体としての意義は、統計学の原理から乖離した分析が数多く行われている現在のコーパス言語学の現状に対し、再考を促すための一石を投じた点にある。本研究の結論である、「①個体（文書や学習者）を観察単位とした分析、②分布図を地図とした分析、③かく乱要因に留意した分析、を行えば、統計学的にも言語学的にも有効な分析ができる」という考え方は、統計学ではごく基礎的な考え方であり、何ら目新しいものではない。本研究の新しさは、これまでの多くのコーパス研究が、このような統計学の原理に基づいて行われてこなかったことを指摘し、これに替わる統計学の原理に基づいた分析法を体系的に提案したところにある。

これまで、文書を観察単位と考えた分析法がほとんど行われてこなかった理由は、数百万語規模のコーパスでは、この分析法で有効に分析できる対象がごく高頻度の言語現象に限られたためであろう。興味深い言語現象を小規模なコーパスで調査するためには、文字、単語、文といった言語単位を観察単位としないかぎり、個体数が少なすぎて分析できない場合が多かったと思われる。このような分析法が、統計の原理に反していることは事実であるが、かといってこれまでの研究で得られた結論の全てが、無意味な結論になっているとは限らない。しかし、その結論が正しいと判断できる根拠もない。正しいか正しくないかは、大規模なコーパスを使用し、統計学的に有効な分析を行って、再度検証してみるしかない。

数億～数百億語規模のコーパスなどは、はるか遠い夢物語であった数十年前であれば、本研究の意義も単なる建前論を述べたものとして、実用性が低かったかも知れない。しかし、現代は、コーパスの規模が次々と大規模化している時代であり、文書を観察単位としても、分析できる言語現象が格段に多くなっている。いつまでも根本的な議論をあいまいにしたまま、統計学の原理から乖離した分析を行う時代は過ぎつつある。それどころか現代では、コーパスが超大規模化した場合、用例数が多すぎて分析が困難になることを懸念する声が上がっている。本研究で提案した分布図を地図として利用する分析法であれば、規模が巨大化しても、詳細に観察する価値が高い文書を特定できるため、せっかく集めた大量のサンプルを再サンプリングして数を減らすなどの措置を取らな

くても、巨大なコーパスが分析しやすい。本研究は、どんどん巨大化していく今後のコーパス分析にとっても、大きな意義を持つと思われる。

本研究の分析法は、学習者コーパスの分析でも、大きな貢献を果たすことができる。学習者コーパスを使用した分析では、そもそも高頻度語を対象として分析を行うことが多いため、現在の規模でも、十分に有効な分析ができるからである。学習者コーパスの最大の問題点は、学習者を無作為抽出して大規模なコーパスを製作することが難しい点にある。このため、学習者コーパスは何らかの偏りを持つことが、ある程度宿命となっている。しかし、これまでの分析法では、コーパスのどこに偏りがあるのかを明らかにすることが難しかった。たとえばKYコーパスでは、各レベルに必要なタスクの種類が決まっているため、言語形式の頻度をレベルごとに比較することには意味がないという注意喚起がなされていた（鎌田，1999:234-6）。しかし、このような注意喚起に従って、あるレベルに特定の偏りが生じていることを明らかにした計量的な研究は、管見のかぎり存在しない。本研究の分析法であれば、このような偏りが特定しやすい。図5.9と図5.10は、KYコーパスとI-JASの学習者レベル別「たら」頻度を比較したグラフである。

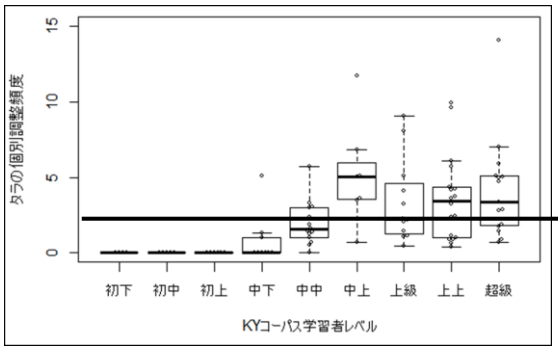


図5.9　KYコーパスの「たら」頻度（再掲）

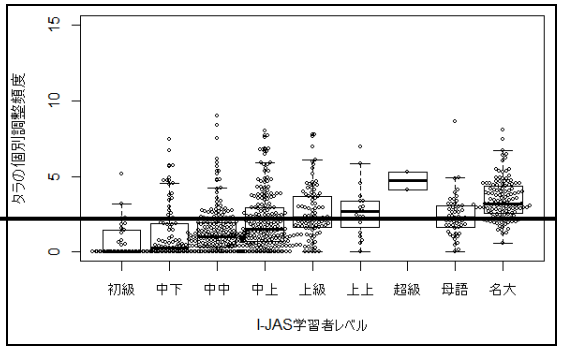


図5.10　I-JASの「たら」頻度（再掲）

この図では、KY コーパスでは中級上の全体が、I-JAS では超級の全体が、それ以外の学習者レベルの傾向とは異なり、レベルの全体が偏っていることが明瞭に観察できる。

また、学習者コーパスの分析法では、これまで使用されてきた調整頻度に替わり、分布観察に基づいた中央値を代表値として使用した方がよい点を明らかにした意義も大きい。学習者の言語使用は、母語話者とは異なって、大きなばらつきを持っているため、調整頻度を使用しても、正確な分析は行いにくい。本研究は、このような問題点に気づかないまま行われてきた学習者コーパスの分析を、大きく改善することができる。

以上、本研究の全体的な意義をまとめれば、これまでの多くのコーパス研究が統計学

の基本から乖離して行われてきた点を指摘し、統計学の基礎に則った、統計学的にも言語学的にも有効な分析法を体系的に提案したところにある。

今後の課題としては、まず、本研究の検証があげられる。本研究の結果、先行研究には必ずしも有効な分析になっていない分析結果が含まれている可能性が示唆された。先行研究を再分析することは、後ろ向きの分析のようでもあり、データの入手等の問題もあって簡単ではないが、本研究の主張が正しいか否かを検証するためにも、一定量の追試を行っていく必要がある。一番望ましいのは、本研究で示した分析法の重要性を多くの研究者に理解してもらい、研究者自らの手で、自分の研究を再分析してもらうことである。

もう一つの課題は、調査対象と分析方法を拡大させることである。本研究が対象とした分析法は、ごく基礎的な方法に留まる。対象とした言語現象も高頻度の単語が多い。同様の分析法であっても、もっと低頻度の言語現象の場合、どれぐらい有効な分析ができるのか、同じ回帰分析でも、マルチレベル分析や多母集団分析のようなより高度な分析はどのように行っていけばよいのかなど、調査対象と分析法の幅を広げて研究していく必要がある。本研究で提案した分析法の有効性を検証し、分析の対象と方法を広げていくことが、今後の課題である。

最後に、コーパスの製作方法に対して、3 点ほど提言しておきたい。その一つは、BCCWJ の出版書籍と図書館書籍の 1 文書当たりの量を増やして、バージョンアップさせることである。BCCWJ の設計思想では、語彙の偏りを防ぐために、固定長の語数が少なく抑えられていた。しかし、本研究の研究結果からすると、もっと長い文書の方が母集団のよりよい標本となった可能性がある。著作権の許可を受け直すなどの問題もあるだろうが、1 文書当たり数万語の固定長ができれば、大きな精度アップが期待できると思われる。

二つ目は、Web コーパスなどの大規模コーパスに、個体の情報を付与することである。本研究の考え方に則れば、言語使用の差異を生み出しているのは個体であるため、個体の情報がないと、言語学的な意味を明確にした分析が難しい Web コーパスを使用した最近の研究では、ドメイン名でデータを集約して比較する研究などが散見されるが、これは、個体情報がないための苦肉の策だと思われる。Web コーパスでは、重複するデータを除くなどの加工が施されているため、厳密には URL を個体と認定できないかも知れないが、そのような細部にこだわるより、URL の情報によって、言語使用の差異を説明できることの方がはるかに重要だと思われる。

三つ目は、学習者コーパスを作るに当たって、データ採取を行う機関を、同一母語に対して少なくとも十か所程度以上に増やすことである。たとえばI-JASの場合、意欲的に多くの種類の母語の調査が行われているが、その調査機関の多くは1機関で行われた調査であり、母語別の調査機関による分布が観測できない。このため、母語別に何らかの差異が観測できたとしても、それが母語の違いによるのか、調査機関の違いによるのかが分からない。調査対象が1母語当たり10機関程度あれば、その分布から、その母語の中心的な傾向が推定できる。全ての言語に対してデータを増やすことは難しいかも知れないが、現在も複数の機関からデータが集められている中国語、韓国語、英語などだけでも、調査機関を増やす整備が進めば、精度の高い母語別の研究が行えると思われる。

使用データ

本研究はコーパスとして『現代日本語書き言葉均衡コーパス』Web 版、および DVD 版（Version1.1）、『タグ付き KY コーパス』（http: //jhlee.sakura.ne.jp/kyc/）、『KY コーパス version1.2』、『多言語母語の日本語学習者横断コーパス：I-JAS』、『名大会話コーパス』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」（https://chunagon.ninjal.ac.jp/ijas/search）を利用した。

また、これらの関連資料として公開されている以下の資料を使用した。

『多言語母語の日本語学習者横断コーパス』「中納言」バージョン 2.4 2.4.2　短単位データ 20180502 版　I-JAS 語数表_20180502 版、『現代日本語書き言葉均衡コーパス』短単位語彙表 ver.1.0、『現代日本語書き言葉均衡コーパス』長単位語彙表 ver1.0、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）.

文献

赤野一郎・堀正広・投野由紀夫（2014）『英語教師のためのコーパス活用ガイド』大修館書店.

Baroni, M. and S. Evert. (2009) “Statistical methods for corpus exploitation.” A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics: An International Handbook Vol. 2*. W. de Gruyter (New York), pp.777-803.

Biber, D. (1993) “Representativeness in cCorpus Design.” *Literary and Linguistic Computing*. 8(4), pp.243-257.

バイバー，D. ・コンラッド，S. ・レッペン，R. （2003）齊藤俊雄・朝尾幸次郎・山崎俊次・新井洋一ほか（共訳）『コーパス言語学―言語構造と用法の研究―』南雲堂.

趙海城（2015）「上級～超級日本語学習者の作文から見た言語産出実態」『第 7 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.293-302.

Church, K, W. (2000) “Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p2.”*Proceedings of COLING 2000.* Saarbrücken, pp.180‐186. (Saarbrücken, Germany). (http://acl.ldc.upenn.edu/C/C00/C00‐1027.pdf)

Dewey, J. (1894) “The theory of emotion I : Emotional attitude." *Psychological Review.* 1(6), pp.553-569.

Evert, S. (2006) “How Random is a Corpus? The Library Metaphor.” *Zeitschrift für Anglistik und Amerikanistik.* 54(2), pp.177-190.

Francis, W. N. and Kučera, H. (1979) BROWN CORPUS MANUAL. MANUAL OF INFORMATION to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.　<http://www.hit.uib.no/icame/brown/bcm.html>　2018.06.21 閲覧.

後藤斉（1993）「「神話」の比喩的用法について-コーパス言語学からのアプローチ」『東北大学言語学論集』2,pp.1-16.
後藤斉（1995）「言語研究のデータとしてのコーパスの概念について　―日本語のコーパス言学のために―」『東北大学言語学論集』4,pp. 71-87.
後藤斉（1997）「コーパスの類型論」『東北大学言語学論集』6 ,pp.27-33.
後藤斉（2003）「言語理論と言語資料　―コーパスとコーパス以外のデータ―」『日本語学』22 (5) ,pp.6-15.
後藤斉（2007）「コーパス言語学と日本語研究」『日本語科学』22, 国立国語研究所, pp.47-58.
Gries, S. (2010) "Useful statistics for corpus linguistics," Aquilino Sanchez Moises Almela (eds.), *A mosaic of corpus linguistics: selected approaches.* Peter Lang (Frankfurt), pp.269-291.
グレンジャー，S.（編）（2008）船城道雄・望月通子（監訳）『英語学習者コーパス入門 ―SLAとコーパス言語学の出会い―』研究社.
樋口耕一（2014）『社会調査のための計量テキスト分析―内容分析の継承と発展を目指して』ナカニシヤ出版.
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.
市原清志・岩本美江子（2006）『カラーイメージで学ぶ統計学の基礎』日本教育研究センター.
庵功雄（2017）『一歩進んだ日本語文法の教え方 1』くろしお出版.
石田基広（2017）『R によるテキストマイニング入門（第 2 版）』森北出版.
石田基広・小林雄一郎（2013）『R で学ぶテキストマイニング』ひつじ書房.
石田将吾・佐藤理史（2010）「エッセイコーパスを用いた日本語テキストの著者推定」研究報告自然言語処理,.2010-NL-198. (6),pp.1-6.
石川慎一郎（2008）『英語コーパスと言語教育―データとしてのテクスト』大修館書店.
石川慎一郎（2012）『ベーシック　コーパス言語学』ひつじ書房.
石川慎一郎（2017）『ベーシック応用言語学　L2 の習得・処理・学習・教授・評価』ひつじ書房.
石川慎一郎・前田忠彦・山崎誠（編）（2010）『言語研究のための統計入門』くろしお出版.
伊藤雅光（2002）『計量言語学入門』大修館書店.
伊藤雅光（2005）「計量言語学とコーパス言語学」『計量国語学』25 (2) ,pp.89-97.
樺島忠夫（1954）「現代文における品詞の比率とその増減の要因について」『国語学』(18), pp.15-20.
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.
樺島忠夫（1963）『表現論－ことばと言語行動』綜芸舎.
樺島忠夫（2009）「語彙量の実態」計量国語学会（編集）『計量国語学事典』朝倉書店, pp.93-97.
樺島忠夫・寿岳章子（1965）『文体の科学』綜芸舎.

鎌田修（1999）「KY コーパスと第二言語としての日本語の習得研究」『第 2 言語としての日本語の習得に関する総合研究』科研研究報告書 08308019 代表者カッケンブッシュ寛子, pp. 227-237.

鎌田修（2006）「KY コーパスと日本語教育研究」『日本語教育』130,pp.42-51.

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』4 (1),pp.43-53.

柏野和佳子・丸山岳彦・稲益佐知子・田中弥生・秋元祐哉・佐野大樹・大矢内夢子・山崎誠（2009）『『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例』特定領域研究「日本語コーパス」平成 20 年度研究成果報告書（JC-D-08-01）.

Kilgarriff, A. (1997) “Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora.” *Proceedings 5th ACL workshop on very large corpora*. pp.231-245.

Kilgarriff, A. (2005) “Language is never, ever, ever, random.” *Corpus Linguistics and Linguistic Theory*. 1 (2), pp.263-275.

木村睦子（1982）「語彙の計量」，佐藤喜代治（編）『講座日本語の語彙』，明治書院, pp.225-243.

金明哲 (2002)「助詞の分布における書き手の特徴に関する計量分析」『社会情報』11 (2), pp.15-23.

金明哲（2009）『テキストデータの統計科学入門』岩波書店.

北居明（2014）『学習を促す組織文化―マルチレベル・アプローチによる実証分析』有斐閣.

小林雄一郎（2017a）『R によるやさしいテキストマイニング』オーム社.

小林雄一郎（2017b）『R によるやさしいテキストマイニング［機械学習偏］』オーム社.

小島寛之（2006）『完全独習　統計学入門』ダイヤモンド社.

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/　(BCCWJ_LB_Stylistics-1.0.zip)　>.

国立国語研究所コーパス開発センター（2015）「『現代日本語書き言葉均衡コーパス』利用の手引き第 1.1 版」国立国語研究所.

栗田佳代子（1996）「観測値の独立性の仮定からの逸脱が検定の検定力に及ぼす影響」『教育心理学研究』44 (2),pp.234-242.

許夏珮（1997）「中・上級台湾人日本語学習者による『テイル』の習得に関する横断研究」『日本語教育』95, pp.37-48.

許夏珮（2000）「自然発話における日本語学習者による『テイル』の習得研究 :OPI データの分析結果から」『日本語教育』104, pp.20-29.

Leech, G.（2008）「序文」グレンジャー，S.（編）（2008）船城道雄・望月通子（訳）『英語学習者コーパス入門 ―SLA とコーパス言語学の出会い―』研究社, pp. xi -xviii.

李在鎬（2009）「タグ付き日本語学習者コーパスの開発」『計量国語学』27 (2) ,pp.60-72.

李在鎬・淺尾仁彦・濱野寛子・佐野香織・井佐原均（2008）「タグ付き日本語学習者コーパスの開発」『言語処理学会第 14 回年次発表大会発表論文集』pp.658-661.

李在鎬・小林典子・今井新悟・酒井たか子・迫田久美子（2015）「テスト分析に基づく「SPOT」と「J-CAT」の比較」『第二言語としての日本語の習得研究』18,pp.53-69.

間淵洋子（2011）「第 5 章　コーパスを利用した研究例」荻野綱男・田野村忠温（編）『講座 IT と日本語研究 5　コーパスの作成と活用』明治書院, pp.165-230.

前川喜久雄（2013）「第 1 章　コーパスの存在意義」前川喜久雄（編）『講座日本語コーパス　1. コーパス入門』朝倉書店, pp.1-31.

丸山直子（2015）「コーパスにおける格助詞の使用実態―BCCWJ・CSJ に見られる分布」『計量国語学』30 (3),pp.127-144.

丸山岳彦・秋元祐哉（2007）『『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 －現代日本語書き言葉の文字数調査－』, 特定領域研究「日本語コーパス」平成 18 年度研究成果報告書（JC-D-06-02）.

丸山岳彦・秋元祐哉（2008）『『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法（2） －コーパスの設計とサンプルの無作為抽出法－』, 特定領域研究「日本語コーパス」平成 19 年度研究成果報告書（JC-D-07-01）.

丸山岳彦・柏野和佳子（2014）「第 2 章　サンプリング」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.22-44.

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011a）『『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用』, 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-01）.

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011b）『『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装』, 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-10-02）.

松田真希子・宮永愛子・庵功雄（2013）「超級日本語話者の談話特性―テキストマイニングを用いた分析―」『国立国語研究所論集』5,pp.43‐63.

マケナリー，T. ・ハーディー，A. （2014）石川慎一郎（訳）『概説コーパス言語学－手法・理論・実践』ひつじ書房.

水本篤・竹内理（2008）「研究論文における効果量の報告のために―基礎的概念と注意点―」『英語教育研究』31,pp.57-66.

水本篤・竹内理（2011）「効果量と検定力分析入門―統計的検定を正しく使うために―」『より良い外国語教育のための方法―外国教育メディア学会（LET）関西支部メソドロジー研究会 2010 年度報告論集―』pp.47-73.

望月通子（2012）「日本語教育における学習者コーパスの構築と ICLEAJ 」『関西大学外国語学部紀要』7,pp.111-119.

森篤嗣（2011）「日本語教育文法のための研究手法」森篤嗣・庵功雄（編）『日本語教育文法のための多様なアプローチ』ひつじ書房, pp.13-55.
森敏昭・吉田寿夫（編著）（2013）『心理学のためのデータ解析テクニカルブック』北大路書房.
森秀明（2016）「BCCWJ 書籍サブコーパスにおける全データ混合分析の検討」『言語処理学会第22回年次大会発表論文集』pp.274-277.
森秀明（2017）「KY コーパスを使用した計量的分析法の現状と課題」『文化』81 (1/2), 東北大文学会, pp. 75-95.
森幸雄（1987）「生態学的データ利用における誤謬の問題 ―ロビンソンの生態学的誤謬問題を中心として―」『Sociologica』12 (1) ,pp.23-38.
村上征勝（2002）『行動計量学シリーズ 6　真贋の科学―計量文献学入門―』朝倉書店.
村田年（2000）「多変量解析による文章の所属ジャンルの判別―論理展開を支える接続語句・助詞相当句を指標として―」『統計数理』48 (2) ,pp.311‐326.
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）.
Oakes, M, P. (1998) *Statistics for Corpus Linguistics*. Edinburgh University Press (United Kingdom).
尾畑伸明（2014）『クロスセクショナル統計シリーズ 1 数理統計学の基礎』共立出版.
小田利勝（2009）『社会調査法の基礎』プレアデス出版.
小木曽智信（2014）「第 5 章　形態素解析」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. 89-115.
小木曽智信・中村壮範（2011）『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版』, 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-U-10-01）.
小椋秀樹（2014）「第 4 章　形態論情報」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.68-88.
小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）「『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 （上）（下）」, 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-10-05-01,JC-D-10-05-02）.
大野晋（1956）「基本語彙に関する二三の研究」『国語学』24,pp.34-46.
Robinson, W. S. (1950) “Ecological Correlations and the Behavior of Individuals.” *American Sociological Review.* 15 (3), pp.351‐357.
迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子（2016）「多言語母語の日本語学習者横断コーパス　International Corpus of Japanese as a Second Language」『国語研プロジェクトレビュー』6 (3) ,pp.93-110.

迫田久美子（編）（2016）『海外連携による日本語学習者コーパスの構築―研究と構築の有機的な繋がりに基づいて― I-JAS 構築に関する最終報告書(International Corpus of Japanese As a Second Language)』科研研究報告書 24251010 代表者迫田久美子，国立国語研究所.

清水裕士（2014）『個人と集団のマルチレベル分析』ナカニシヤ出版.

白井匡人・三浦孝夫（2012）「確率モデルによる品詞分類の特徴推定」『第 4 回データ工学と情報マネジメントに関するフォーラム 論文集（第 10 回日本データベース学会年次大会）』pp.(n, d.) http：//db-event.jpn.org/deim2012/proceedings/detail.html#e10　E10-3.2017.07.29 閲覧.

砂川有里子（2011）「日本語教育へのコーパスの活用に向けて」『日本語教育』150, pp.4-18.

スタッブズ，M.（2006）南出康世・石川慎一郎（監訳）『コーパス語彙意味論―語から句へ』研究社.

高田智和・小林正行・間淵洋子・大島一・西部みちる・山口昌也（2009）『JIS X 0213:2004 運用の検証』，特定領域研究「日本語コーパス」平成 21 年度研究成果報告書（JC-D-09-01）.

田中真理（1999）「OPI における日本語ヴォイスの習得状況：英語・韓国語・中国語の場合」『第 2 言語としての日本語の習得に関する総合研究』科研研究報告書 08308019 代表者カッケンブッシュ寛子, pp. 335-350.

田野村忠温（2014）「第 6 章　BCCWJ の資料的特性―コーパス理解の重要性―」田野村忠温（編）『講座日本語コーパス　6．コーパスと日本語学』朝倉書店, pp.119-151.

投野由紀夫・金子朝子・杉浦正利・和泉絵美（2013）『英語学習者コーパス活用ハンドブック』大修館書店.

豊田秀樹（編著）（2009）『検定力分析入門―R で学ぶ最新データ解析―』東京図書.

坪根由香里（2002）「OPI における韓国語話者の『の』の使用と習得」『小出記念日本語教育研究会論文集』10,pp.55-70.

Upton, G. and Cook, I.（2011）白旗慎吾（監訳）『統計学辞典』共立出版.

Woods, A., Fletcher, P., Hughes, A.（1986）*Statistics in Language Studies*. Cambridge Textbooks in Linguistics. Cambridge University Press (New York).

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）.

山内博之（1999）「OPI 及び KY コーパスについて」『第 2 言語としての日本語の習得に関する総合研究』科研研究報告書 08308019 代表者カッケンブッシュ寛子, pp. 238-245.

山内博之（2009）『プロフィシエンシーから見た日本語教育文法』ひつじ書房.

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.

山崎誠（2016）「語彙の量的構成」斎藤倫明（編）『講座　言語研究の革新と継承Ⅰ　日本語語彙論Ⅰ』ひつじ書房, pp.105-133.
山崎誠（編）（2014）『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.
山崎誠・前川喜久雄（2014）「第 1 章　コーパスの設計」山崎誠（編）（2014）『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.1-21.
安本美典（1960）『文章心理学の新領域』東京創元社.
安本美典（1965）『文章心理学入門』誠信書房.
吉田寿夫（2001）『本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本』北大路書房.

本論文に関する外部発表一覧

第 2 章　先行研究

森秀明「KY コーパスを使用した計量的分析法の現状と課題」『文化』第 81 巻第 1・2 号，pp.75-95，2017 年。

第 4 章　分布観察の方法

森秀明「BCCWJ 図書館サブコーパスの代表性試論」『第 8 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.19-28，2015 年 9 月 1 日。

森秀明「BCCWJ を使用した格助詞頻度の分析法―図書館サブコーパスに出現した格助詞ノを中心に―」『言語科学論集』20 号, 東北大学大学院文学研究科言語科学専攻, pp.119-130，2016 年。

第 5 章　代表値と分布図を併用した頻度比較の方法

森秀明「I-JAS と KY コーパスにおける量的な性質の比較」『2017 年度日本語教育学会支部集会予稿集【東北支部】』，pp.21-26，2017 年 12 月 10 日。

森秀明「学習者コーパスを使用したレベル別頻度比較の方法」『Learner Corpus Studies in Asia and the World』vol.3，pp. 303-322，2018 年 3 月。

第 6 章　カイ二乗検定の方法

森秀明「コーパス間における単語使用率の比較―観察単位（ケース）は単語か文書か―」『計量国語学』31 巻 3 号，pp. 205-221，2017 年。

第 7 章 回帰分析の方法

森秀明「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.9-22，2017 年 9 月 5 日。

森秀明「連体助詞の「ノ」と文体の関係」『言語資源活用ワークショップ 2018 発表論文集』，　国立国語研究所，pp.34-46，2018 年 9 月 4 日。

森秀明「コーパス分析における生態学的誤謬」『計量国語学会第六十二回大会予稿集』pp.19-24，2018 年 9 月 29 日。

謝辞

本研究を進めていく中で、多くの方々にお世話になりました。東北大学日本語教育学研究室の才田いずみ先生、田中重人先生、小河原義朗先生、島崎薫先生には、貴重なご意見とご指導を賜りました。心から感謝申し上げます。特に才田先生と田中先生には、修士課程入学以来、8 年間の長きに渡り、さまざまな場面でご指導を賜りました。才田先生には、いつもやる気と刺激をいただきました。先生には海外で研修や発表を行う機会を何度もいただき、その何回かでご一緒させていただくことができたことは、望外の喜びでした。田中先生には、統計学の基本と楽しさを教わりました。この研究のきっかけは、田中先生の授業で使わせていただいた社会調査のデータでは、色々な多変量解析ができるのに、どうしてコーパスの分析ではそれができないのだろうと思ったことにあります。観察単位、生態学的誤謬なども先生の受け売りで、先生のご指導なしには、この研究はなしえませんでした。また、現在は他校に移られましたが、名嶋義直先生、助川泰彦先生にもご指導を賜りました。

東北大学言語学研究室の後藤斉先生には、コーパス言語学の基本を教わりました。幾度となく論文をお読みくださり、ともすると計量的な分析に走りがちになる筆者に、言語学的な視点から考える大切さを教えてくださいました。また、国語学研究室の大木一夫先生、小林隆先生、言語学研究室の木山幸子先生にもご指導を賜りました。心から感謝申し上げます。

研究室の先輩や後輩の皆様からも貴重なアドバイスや励ましをいただきました。特に、年末の忙しい時期に論文のチェックをしてくださった佐竹華奈さん、于凌越さんには心から感謝申し上げます。

最後に、家族にも感謝の言葉を述べたいと思います。社会人である私が、長い時間をかけてこのように研究できたのも、家族の寛容と支えのおかげです。本当にありがとうございました。

多くの皆様のお蔭で博士論文を完成することができました。ここに御礼を申し上げられなかった方々も含め、これまでお世話になった全ての方々に、改めて謝意を表します。

森　秀明