BCCWJ 図書館サブコーパスの代表性試論

森　秀明（東北大学大学院文学研究科）†

"BCCWJ Library Sub Corpus" And Its Representativeness

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

要旨

『現代日本語書き言葉均衡コーパス』（BCCWJ）の中で、統計分析に適するのは固定長データだと言われている。しかし固定長データのサイズはそれほど大きくない。一方、Sinclair（1991）、バイバー、コンラッド、レッペン（2003）などにより、サイズが小さいコーパスの代表性はさほど高くないことが主張されている。BCCWJ のマニュアルには、語彙の偏りを防ぐためにサンプルを短くしたとの記述が見られるが、その効果を具体的に検討した報告書類は見いだせない。このため語彙表を使用して固定長と可変長の頻度比較による検証を行った。この結果、高頻度語はデータ量に正比例して頻度が増加するが、低頻度語や特定のトピックに使用されやすい固有名詞と普通名詞などは、頻度がばらついて増加することが分かった。代表性が高ければ基本的に頻度のばらつきは生じないと考えられるため、これらの代表性はそれほど高くない可能性がある。

１．研究の目的

あるコーパスが、推定対象の言語を正確に反映していることを代表性と言う。『現代日本語書き言葉均衡コーパス』の「図書館サブコーパス」（以下 BCCWJ の図書館 SC のように表記する）は、都内公立図書館の蔵書を現実母集団とし、そこからデータを無作為抽出して製作されたコーパスであり、高い代表性を持つと考えられている。しかし田野村 (2014) など一部の研究を除けば、その代表性を検討した研究は少ない。

あるコーパスがどれほどの代表性を持つかを実証することは難しい。図書館 SC の場合、現実母集団の蔵書約 33.5 万冊の全文コーパスを作り、それと比較すれば実証できるわけだが、全文コーパスを作るのが現実的に難しいからこそサンプリングコーパスを作っているという関係になっている。このため代表性の検証は、コーパスの設計方針を検討したり、他のコーパスによる検索結果の比較を行うなどの傍証を積み重ねていくしかないと考えられる。ここでは主に設計方針の検討と語彙表の観察から図書館 SC の代表性を検証する。

以下、第 2 節では図書館 SC の設計方針を検討する。第 3 節では語彙表を概観する。第 4 節では固定長の単語の頻度が可変長で何倍になっているのかを中心に調査する。最後に第 5 節でまとめを述べる。

２．設計方針の検討

コーパスの設計で特に重要な点は、どのような方法でサンプルを抽出するかという点と、サンプルの数×サンプルの長さ＝コーパスのサイズをどれぐらいの大きさにするかという 2 点だと思われる。ここでは主にサンプルサイズの問題に絞って検討する。

図書館 SC の設計方針を検討するには、類似の方針で製作されたコーパスの設計方針と比較すると、その特徴が明確になる。このため、世界的に代表性が高いと評価されている British National Corpus（以下 BNC と言う）の設計方針を簡単に確認しておく（Burnard (ed.)，2007；

† hideaki@moriharuo.com

アシュトン、バーナード，2004）。

BNC は 1995 年にイギリスで製作されたコーパスで、総語数は約 1 億語である。そのうち書籍データは 1411 冊×平均 3.6 万語＝約 5 千万語となっている。書籍はテキストタイプを情報伝達散文（8 種類）、文芸作品、未分類の計 10 種類に独自に分類し、ベストセラーの一覧リストや図書館の貸し出し冊数を参考に選抜した。さらにそれぞれの書籍から 4 万語を目安にサンプルを取得し、4 万語に満たない書籍は全文を、4 万語以上の書籍は最大 4.5 万語を採用した。この結果、サンプル当たりの語数は平均で約 3.6 万語となっている。このような方法は世界で初めて製作された Brown コーパス（500 冊×2,000 語＝100 万語）などと類似の方法である。

次に BCCWJ の図書館 SC のサンプリング方法を概観する（国立国語研究所, 2011 ; 丸山、柏野，2014）。図書館 SC は、書き言葉の流通の実態に着目し、東京都内の公立図書館で重複所蔵されていた 1986 年～2005 年発行の書籍約 33.5 万冊分、およそ 479 億字を母集団とした。サンプルの選択に当たっては全書籍のページをランダムに並べた長大なリストを作り、これを 20 年間の出版年と日本十進分類法の 11 分類の組み合わせによって 220 層に区分した。そしてそれぞれの層から復元無作為抽出法によって 10,551 箇所を選択した。この箇所に該当した書籍からさらに無作為に場所を選んでサンプルを抽出した。

抽出に当たっては、それぞれのサンプルから記号等を除いた文字数で 1 千字に固定した固定長と、それぞれのサンプルにおける節や章などの文章のまとまりに留意し、最大 1 万字まで抽出した可変長という二種類のデータを抽出した。田野村（2014，p. 112）の表 6.3 によれば、記号等を含めた文字数の固定長平均は 1,170 字、可変長平均は 5,039 字で、可変長の文字数は固定長の約 4.3 倍になっている。語数に直してコーパスサイズを計算すると、固定長は平均 635 語×10,551 サンプル＝約 670 万語、可変長は平均 2,738 語×10,551 サンプル＝約 2,889 万語で、これも約 4.3 倍である。ただし、固定長と可変長は必ずしも重複していないため、この両者を足して重複を除いたデータが最大となる。それをここでは「両方データ」と呼ぶ。両方データのサイズは平均 2,879 語×10,551 サンプル＝約 3,038 万語である。図書館 SC の最大サイズは両方データの約 3 千万語だが、これはサンプルごとの文字数が異なるので均衡ではない。このため BCCWJ のマニュアルには、統計分析に適するのは固定長データであると記されている（国立国語研究所，2011，p. 23）。

図書館 SC は、最大サイズで言えば BNC 書籍データの 6 割あるが、統計分析に適するサイズは 13.4%しかなく、思いのほか小さなコーパスになっている。もし、固定長の文字数を可変長平均の 5 千字にしていたら、統計分析に適するデータで 3 千万語のコーパスが出来上がったはずである。仮に図書館書籍のみで 1 億語のコーパスを作るとしたら、1 サンプルから約 1 万語を抽出すればよい。これならもっと簡単に 1 億語のコーパスが作れたと思われる。様々な選択肢が考えられた中で、なぜ BCCWJ では統計分析に適するとされる固定長の長さを、約 1 千字と言うごく短い長さにしたのであろうか。これを確認するため、BCCWJ の報告書類を閲覧したが、その根拠を実証的に記述した報告は探し当てることができなかった。その代わり、その意図がくみ取れる下記のような文章が散見された。

> 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。例えば、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，2011，p. 1）

> 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にすると，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，p. 26）

これらの記述からすると、固定長の長さを短くしたのは、特定の書籍による語彙の偏りを低減させるためであったことが分かる。しかしこれとは逆に BNC のガイドブックには、語彙の偏りを解消するためにサンプルを長くしたと受け取れる次の記述が見られる。

> Sinclair（1991: 24）は、Brown コーパスと LOB コーパスについて，「この 2 つのコーパスは広い範囲のテキストに出現する比較的頻度の高い単語についてのみ信頼性の高い情報を与えてくれる」と述べています。特定のテキストタイプだけに出現するような単語については，「サンプルが短すぎるのでサンプルのバランスをとるのに必要なサブカテゴリー自体が合理的なサンプルとはなり得ていない」との理由から，「信頼性はそれほど高くない」という評価を下しています。コーパスの規模を大きくし，それぞれのサブカテゴリーにさらに大きなサンプルを収集することで，この問題はいくぶん解決できるでしょう。（アシュトン、バーナード，2004，p. 30）

また、丸山、柏野（2014）が指摘する 1 冊の書籍を丸ごと収録した場合の弊害については、Sinclair（1991）に次の記述が見える。

> The penalties to pay for including whole documents are that in the early stages of gathering, the coverage will not be as good as a collection of small samples and the peculiarities of an individual style or topic may occasionally show through into the generalities. As against these short-term difficulties, there is a positive gain in the study of collocation, which requires very large corpora to secure sufficient evidence for statistical treatment.（Sinclair，1991，p. 19）
> 丸ごとの書籍を収録する弊害は、収集の初期に現れる。この段階のカバー範囲は、小さなサンプルを集積したコーパスと同じぐらい良くないため、一般性より個別のスタイルやトピックによる特殊性がしばしば見られる。このような初期の困難を越えるに従って、コロケーションの研究では、巨大なコーパスでなければ得られないほどの統計的に安定した十分な証拠が得られる。（発表者意訳）

Sinclair は、全文採用のデータを経時的に次々と収集していくモニターコーパスの提唱者である。上記の引用で「収集の初期」のような表現があるのは、モニターコーパスが念頭にあるからだ。しかし、これは時期の問題と言うより収集量の問題と捉えることができる。

モニターコーパスの代表例には Sinclair が監修した Bank of English があるが、これも高い代表性を評価されているコーパスであり、丸山、柏野（2014）が指摘するようなサンプルの全文採用による語彙の偏りは報告されていない。

さらに、コーパスサイズと代表性については、次のような指摘もある。

> LOB Corpus による頻度一覧表によって，コーパスに基づく語彙調査の難題の 1 つも明確になってくる。具体的には，単語の意味と用法を研究するのに，非常に巨大なコーパスが必要になるという点である。つまり，100 万語のコーパスでは，多くの単語に対して，意味のある一般化を行うのに十分なデータを提供できない。頻度数と言うのは，コーパスの非常に頻度の高い単語には比較的信頼性があるが，単語の意味や連語パターンを分析するためには，生起回数が非常に多いものでなければならない。〔・・・・・〕さらに，小さなコーパスの場合，頻度がただ単に中程度の単語を含むか，それとも頻度がまれな単語を含むかどうかは，コーパス内の各テキストに描かれるトピックの違いに大きく左右される。〔・・・・・〕しかしながら，さまざまな多くのテキストを含む非常に大きなコーパスであれば，より広範なトピックが描かれているはずであり，その結果，単語の頻度が個々のテキストによって受ける影響は少なくなる。（バイバー、コンラッド、レッペン，2003，p. 36）

以上の引用からすると、丸山、柏野（2014）が指摘するサンプルを長くすることによる弊害は、確かに収集の規模が小さい場合は懸念されるが、コーパスのサイズを大きくすればその問題は解消し、より高い代表性が得られるとする考え方が存在することになる。図書館 SC の固定長データは、10,280 冊の書籍から 10,551 サンプルを取得しており、トピックの多様性は十分であるように思われるが、サンプル長が平均 635 語とごく短いため、サイズが小さいコーパスになっている。このことによって代表性が十分に高まっていない可能性も考えられる。

３．図書館 SC 語彙表の概観

コーパスのサイズが小さいことで、図書館 SC にはどんな問題が生じるのだろうか。これを確認するため、ここでは「主要コーパス語彙表」と「短単位語彙表データ」を概観する[1]。これらの語彙表はそれぞれに特色が異なる。「主要コーパス語彙表」では語彙の中から機能語が除かれているが、ある単語がいくつのサンプルに出現したかというサンプル頻度が記載されている。ただし可変長や両方データの頻度は載っていない。「短単位語彙表データ」は機能語の頻度と可変長の頻度が記載されているが、サンプル頻度や両方データの単語頻度は載っていない。サンプル頻度は単語の頻度とは質の異なる情報、例えばどれぐらい多くのサンプルに共通して使用されるかで単語の一般性を見るといった情報が得られるため、ここでは両者を併用するが、両者では収録語の対象や語数が異なり、各単語の頻度にも一部に違いが見られるため、以後の分析では必ずしもデータ数が一致しない。

表 2 は、「主要コーパス語彙表」所収の 86,002 語について、単語頻度別に単語数を数えた表、表 3 はサンプル頻度別に単語数を数えた表である。表 2 の単語頻度では、頻度 1 が

[1] これらの語彙表は http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html（国立国語研究所の HP）からダウンロードできる。

25.8%、頻度 2〜5 が 32.0%で、頻度 5 以下で 57.8%になっている。コーパスのサイズが小さいため、頻度が低い単語が大量にある。表 3 のサンプル頻度では、頻度 1 が 36.4%、頻度 2〜5 が 30.3%で、頻度 5 以下で 66.7%である。表 4 は、「短単位語彙表データ」で固定長と可変長が重複する単語 83,232 語について可変長の単語数を数えた表である。このデータには機能語が 166 語加わっているが､固定長と重複した単語で数えると総語数が少なくなる。表 4 を見ると頻度 1 が 7.1%、頻度 2〜5 が 19.3%で、頻度 5 以下で 26.4%、頻度 20 以下で 55.0%となっている（サンプル頻度はデータがないため不明である）。可変長は固定長の 4.3 倍のサイズがあるため、高頻度語の割合が高くなっている。

表 2 固定長の単語頻度

単語頻度	単語数	
1	22201	25.8%
2〜5	27523	32.0%
6〜10	11562	13.4%
11〜20	8996	10.5%
21〜50	7683	8.9%
51〜100	3355	3.9%
101以上	4682	5.4%
合計	86002	100.0%

表 3 固定長のサンプル頻度

サンプル頻度	単語数	
1	31295	36.4%
2〜5	26032	30.3%
6〜10	9427	11.0%
11〜20	7008	8.1%
21〜50	6093	7.1%
51〜100	2601	3.0%
101以上	3546	4.1%
合計	86002	100.0%

表 4 可変長の単語頻度

単語頻度	単語数	
1	5938	7.1%
2〜5	16027	19.3%
6〜10	11317	13.6%
11〜20	12484	15.0%
21〜50	15100	18.1%
51〜100	8357	10.0%
101以上	14009	16.8%
合計	83232	100.0%

これらの表を見ると、コーパスのサイズが小さいことによる最大の問題は、その代表性を云々する以前に、あまりにも頻度の少ない単語が多いことであるのが分かる。国立国語研究所（2011，p. 23）は、統計分析に適するのは固定長であるとしているが、統計分析にはデータの質だけでなくデータの量も重要である。固定長では頻度 5 以下の単語が 6 割弱あり、これらを使用して統計的に有意な分析を行うのは困難だと思われる。それならむしろ文字数のばらつきを考慮に入れながら可変長の単語頻度を使用したり、文字数のばらつきには比較的影響されにくいサンプル頻度を指標にすることを考えてみても良いだろう。分析の対象や方法によっては、可変長（正確には最もサンプル長が長い両方データ）の方が、統計分析に適していることも考えられる。「単語の意味や連語パターンを分析するためには，生起回数が非常に多いものでなければならない。」（バイバー、コンラッド、レッペン，2003，p36）という指摘は、重く受け止める必要があるだろう。

４．固定長頻度と可変長頻度の比較

図書館 SC の固定長データは、サンプル長が短くコーパスサイズが小さいため代表性が十分に高まっていない可能性が考えられる。これを検証するには、どうすれば良いだろうか。大規模な調査が可能なら、固定長データを 100 字ごとに区切ったデータを作り、コーパス文字数の増加に対する全単語の頻度増加率を観察するのが良いと思われる。文字数の増加に対して頻度が一定に増加しているなら代表性は高く、増加率が不安定なら代表性は高くないと考えられる。代表性の高いコーパスとは、どんどんサンプル長やサンプル数を増大させた結果、データ量の増加に対して頻度の増え方が正比例するようになったコーパスのことである。そのような状態に達したコーパスなら、もうそれ以上サンプル長やサンプル数を増やす必要はない。そのコーパスで得られた頻度に一定数をかければ母集団の正確な頻度が推定できる。それに対し字数が増加するたびに頻度の増加率が変わるなら、まだ母

集団を推定する準備が整っていないと言える。これは代表性が低いコーパスである。代表性とは、コーパスが母集団の正確な縮尺になっていることである。しかし、ある単語で例えば固定長の 800 字→900 字段階と 900 字→1 千字段階を比較してまだ増加率に揺れがあるなら、正確な縮尺になり切っていない可能性が高いと考えられる。

ただし、このような検証は相当に大規模な研究になる。これをもっと簡便に行うには、固定長データと可変長データの比較が考えられる。しかし、可変長は個々のサンプルごとに文字数が異なるため、統計分析には適さないとされている。例えば A という単語の頻度を可変長で調べた場合、固定長頻度の 4.3 倍になっていれば正確で、0.1 倍とか 10 倍になっていれば不正確だとは言えないとする考え方もあるだろう。A という単語が短い可変長データにのみ出現する単語であれば 0.1 倍になることもあるし、長い可変長データにのみ出現する単語であれば 10 倍になることもあり得るからである。しかし、現実的には個別の単語が可変長のサンプルの長さに連動した出現傾向を持っているとは考えにくい。機能語のような高頻度語なら、短いサンプルでも長いサンプルでも、その出現傾向はほぼ同じだと思われる。中・低頻度語の場合も、どの単語が短いサンプルに出現し、どの単語が長いサンプルに出現するかは、十分ランダムになっていると考えられる。このため固定長と可変長の比較は、厳密な正確性には欠けるかも知れないが、図書館 SC に出現する語彙の全体像を簡便に観察するための調査としては、ある程度妥当なものだと考えられる。そこでここでは、固定長と可変長の頻度を比較し、その増加率がどれほど安定しているかを調査する。データには「短単位語彙表データ」を使用する。

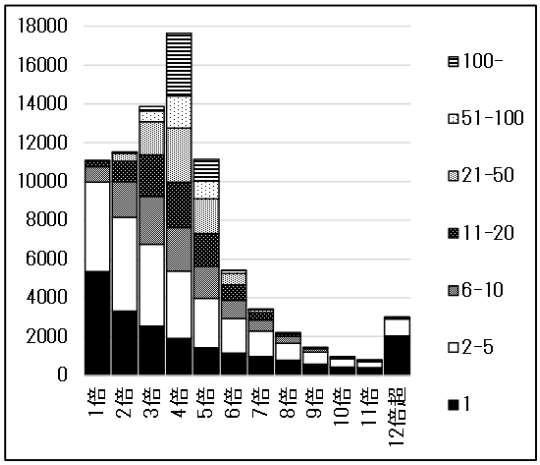

図 2 頻度別・可変長倍率ごとの単語数

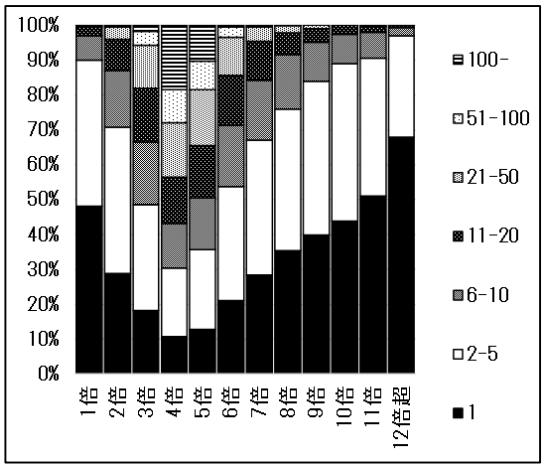

図 3 可変長倍率ごとの単語の頻度割合

図 2 は、表 2～4 の頻度区分ごとに分けた固定長の単語の数を、可変長の頻度倍率ごとに積み上げたグラフである（1 倍は 0.51 倍～1.50 倍の範囲）。この頻度倍率÷4.3 が増加率である。図 2 を見ると固定長の頻度は可変長で 4 倍になっているものが最も多い。つまりデータ量にほぼ正比例して増加している単語が最も多いということが分かる。

図 3 は、図 2 を割合で表したグラフである。高頻度の単語は 4 倍と 5 倍に多く、ここから倍率が離れるに従って低頻度の単語の割合が多くなる。頻度 100 以上の高頻度語は、4 倍が 69.8%、5 倍が 24.5%で、この二つで 94.3%になる。このことから高頻度語の頻度はデータ量の増加にほぼ正比例して増加することが分かる。その一方で低頻度語は、様々な倍率になる。この現象は、低頻度語の不安定さを示すものであり、固定長における低頻度語の

頻度が必ずしも正確だとは言い切れないことを示唆している。現在の固定長データでは頻度 1〜5 になっている単語でも、サンプリングをやり直した別バージョンの固定長データなら、頻度が 1〜15 などのように変わる可能性も考えられる。

この議論を、図 4、5 の箱ひげ図[2]を使用して整理して見よう。図 5 は図 4 の拡大図、表 5 はこれらの記述統計量である。図 4 の横軸は基本的に表 2〜4 の頻度区分と同じもので、1 は 1、2-5 は 5、6-10 は 10 のように区分の最大値で表記している。表 2 と異なり、図 4 では 101-1,000 と、1,001 以上も分けて描いた。10,000 というラベルは、固定長の頻度が 1,001 を超える超高頻度語につけている。

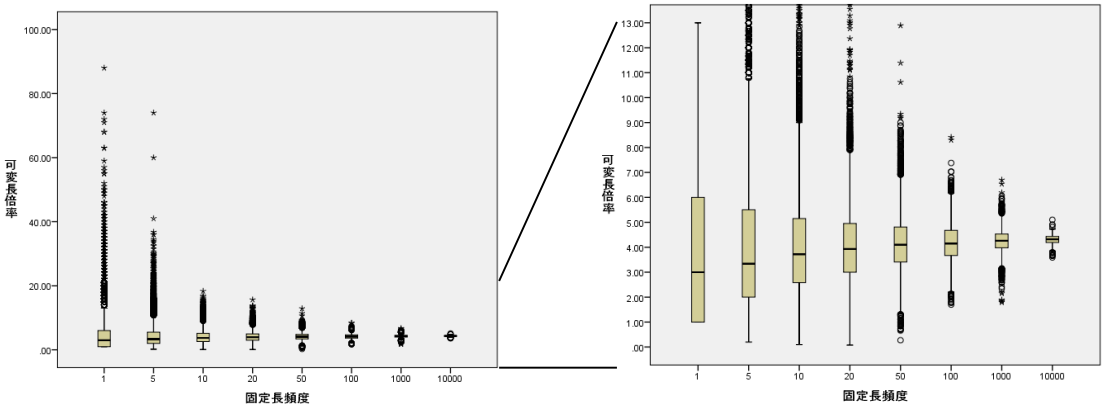

図 4 固定長頻度別可変長倍率分布（全体）　　図 5 固定長頻度別可変長倍率分布（拡大）

表 5　固定長頻度区分別における可変長倍率の記述統計量

	度数	平均	標準偏差	標準誤差	平均値の 95% 信頼区間		最小	最大
					下限	上限		
1	20826	5.0152	5.52557	.03829	4.9402	5.0903	1.00	88.00
5	26576	4.1695	3.32504	.02040	4.1295	4.2095	.20	74.00
10	11371	4.0486	2.09328	.01963	4.0102	4.0871	.10	18.34
20	8925	4.0759	1.61271	.01707	4.0424	4.1093	.08	15.64
50	7510	4.1477	1.12341	.01296	4.1223	4.1731	.27	12.89
100	3358	4.1883	.78040	.01347	4.1619	4.2147	1.70	8.41
1000	4130	4.2502	.47228	.00735	4.2357	4.2646	1.79	6.70
10000	536	4.2961	.20722	.00895	4.2785	4.3137	3.59	5.10
合計	83232	4.3582	3.51331	.01218	4.3343	4.3820	.08	88.00

表 5 で 10,000 の度数を確認するとわずか 536 しかない。これを品詞ごとに高頻度順に示せば、助詞「の・に・て」、動詞「する・いる・ある」、固有名詞「日本・アメリカ・東京」などになる。頻度 1,001 付近の単語は「働く・進む・内容・基本」などである。図 5 を見ると、10,000 の箱ひげ図は、他の箱ひげ図と比べて極めて小さいことが分かる。これはこの群に属する 536 語が可変長のデータでほとんどばらつくことなく、4.3 倍付近に集中していることを表している。表 5 で確認すると平均は 4.296、標準偏差は 0.207 である。具体的な単語で見ると助詞の「の」は固定長頻度の 342,113 が可変長では 1,473,404 と 4.31 倍に、固有名詞の「日本」が 8,846 から 37,131 と 4.20 倍に、動詞の「働く」が 1,001 から 4,397 と

[2] 箱ひげ図は、真ん中の黒い線が中央値、箱の上下が 75 パーセンタイルと 25 パーセンタイル、ひげの上下が 90 パーセンタイルと 10 パーセンタイルの位置を表す。ひげの外の○や☆は外れ値である。

4.39 倍になっている。これらの高頻度語が可変長ではそのデータ倍率とほぼ同じ 4.3 倍になっているのは、これらの頻度が極めて高く、高い代表性を持っているからだと考えられる。図書館書籍の母集団の文字数はおよそ 479 億字であるから、これらの固定長頻度を 4,790 倍にすればほぼ母集団の頻度と同じになると考えて良いだろう。

その一方で 1 の箱ひげ図は、90 パーセンタイルが可変長倍率 13 倍となるなどばらつきが大きい。図 4 を確認すると最大で 88 倍になっている。固定長で頻度 1 の単語が、可変長になると頻度 1 から頻度 88 にまでばらついて増加していることが分かる。これらの頻度を 4,790 倍にしたからと言って、母集団の正確な頻度が推定できるとは思われない。つまり、代表性は高くないと考えられる。なお、図 5 の箱ひげ図で、低頻度になるほど中央値が 3 に近づく現象が観察される。これは低頻度になるほど増加率が低くなる単語が多いためである。固定長で頻度 1 の単語には、可変長になっても頻度が 1 のままである単語も多い。これらの多くは母集団でも頻度 1 のままであることが予想される。その意味では、低頻度語の中にも代表性が高い単語が含まれていることになる。

図書館 SC の低頻度語は、可変長における頻度倍率が大きくばらつくため、その多くの代表性は高くないと考えられる。それでは低頻度語はなぜこれほどまでばらつくのであろうか。次にこの問題を調査する。

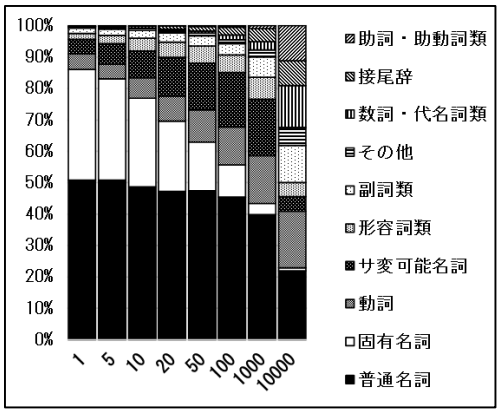

図 6 固定長頻度別品詞割合

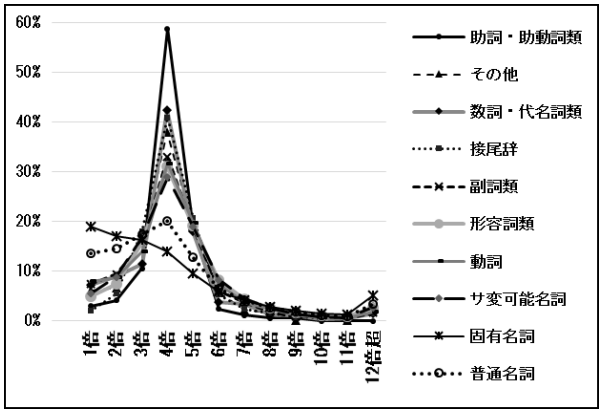

図 7 品詞ごとの可変長倍率分布

図 6 は、表 5 の頻度区分ごとに固定長データの品詞割合を示したグラフである。これを見ると低頻度語の大半は普通名詞と固有名詞であることが分かる。普通名詞は頻度区分が 1000 の場合でも 4 割程度を保つが、固有名詞は頻度区分が上がるにつれてその数を激減させる。この理由は、固有名詞の多くが特定のテキストにしか出現しない特定の単語であるためだと思われる。図 7 は、各品詞ごとに可変長で何倍になりやすいかを表したグラフである。最も高頻度語である助詞・助動詞類ではその 6 割が 4 倍、9 割以上が 3〜5 倍の範囲である。これに比べ、普通名詞と固有名詞はその多くが 1〜6 倍に散らばっている。グラフが見にくくて恐縮だが、固有名詞は 12 倍超の割合も 5%以上ある。

この二つのグラフから分かることは、固有名詞や普通名詞には低頻度の単語が多いこと、固有名詞や普通名詞は可変長になると様々な倍率で増加するということである。図 6 の普通名詞は大半の頻度区分で 5 割弱を維持するが、この普通名詞の内部でも一部のテキストでしか使われない特定の単語と多くのテキストで使われる一般的な単語の交替現象が起きていると考えられる。つまり低頻度語が大きくばらつく理由は、品詞の特性による影響、

すなわち特定のテキストに出現する特定の単語の出現パターンが原因である可能性が高い。

これを具体的な単語で観察してみよう。表 6 は「トマト」という普通名詞がどのサンプルに何個出現したかを数えた表である。固定長の頻度が多いものから順に 8 サンプルを表示している。固定長ではこの他に 66 サンプルに出現し、全体合計は 201 である。このうち上位 8 サンプルで 89 と全体の 44.2%に達するため、「トマト」の頻度ではこれら 8 サンプルの影響が強いことが分かる。書名を見ると料理関係や野菜作りのトピックが多く、「トマト」という単語は特定のトピックで多用される単語であることが確認できる。

問題は、このような単語がうまくサンプリングできているかどうかである。図 8 は、それぞれのサンプルのどの位置に「トマト」という単語が出現するのかを表している。縦軸は表 6 のNo.に対応し、整数の位置に固定長と可変長を含めた全体（両方データ）を、整数＋0.5 の位置に固定長の出現状態をプロットしている。両方データの表示にある×は、サンプルの末尾を表している。横軸は語数で、目盛りは記号等を含む固定長平均の 750 語で区切ってある。

表 6 サンプル別「トマト」の出現数

NO.	書名	固定長	可変長	倍率
8	ほんじょの虫干。	6	6	1
7	トマト弁護士被告人の甘い囁き	7	7	1
6	永田農法・驚異の野菜づくり	7	36	5.2
5	知っておきたいキッチンハーブ	10	21	2.1
4	ケンタロウの野菜がうまいッ！	10	28	2.8
3	シニアのためのライトフレンチ	14	10	0.8
2	わかりやすいイタリア料理	16	0	0
1	食べるのが大好き	19	21	1.2
	小計	89	129	1.5
	その他(固定長66冊、可変長160冊)	112	415	3.8
	合計	201	544	2.8

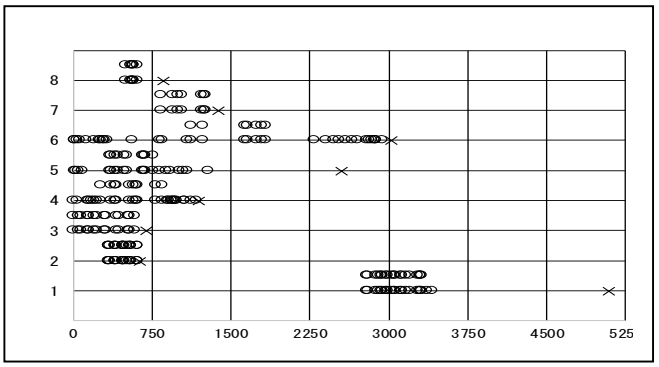

図 8 「トマト」の出現位置（上：固定長・下：全体）

No.1『食べるのが大好き』では両方データの語数は 5080 語で、そのうち真ん中から後半で集中的に「トマト」が出現する。両方データで見れば、「トマト」が出現しているのはサンプルの 1/7 に過ぎないが、固定長のサンプル長は短いため、全体に万遍なく出現していることが分かる。No.5『知っておきたいキッチンハーブ』でも、両方データでは後半には 1 語も出現しないが、固定長は前半の「トマト」が頻出する部分のみを抽出しているため、サンプル全体の平均的な頻度より多くなっている。同様の問題はNo.7『トマト弁護士被告人の甘い囁き』でも見られる。No.2、3、4、8 は両方データ自体が短いため、一見問題があるようには見えないが、サンプルを長くした場合、「トマト」と言う単語が残りの部分には全く出現しない可能性も否定できない。これらのサンプリング状況を見ると、固定長データから母集団の「トマト」の頻度を推定すれば、その頻度をかなり過大評価することになるのではないかと思われる。この理由は固定長の抽出範囲が短すぎて、テキスト全体における出現確率を正確に反映できていないためである。BCCWJ の設計方針はサンプルを無作為抽出することで各サンプルの標本誤差が均衡化されることを期待するものだが、そのような大数の法則は大量のデータでしか働かない。サンプル頻度が少ない場合は個々のサンプルが個々のテキストをある程度正確に反映している必要があると考えられる。

「トマト」は固定長のランクで 2689 位、可変長で 3862 位の高頻度語である。固有名詞や一部の普通名詞は特定のテキストに出現しやすいだけでなく、その出現の仕方も一か所に固まって出現しやすいなど特殊であるため、単語頻度 201、サンプル頻度 74 の高頻度語であっても、短いサンプル長で正確なサンプリングを行うのは困難なのだと思われる。

５．まとめ

『現代日本語書き言葉均衡コーパス』（BCCWJ）の中で、統計分析に適すると言われているのは固定長データである。しかしこれらのサイズは思いのほか小さい。一方、Sinclair（1991）、バイバー、コンラッド、レッペン（2003）などにより、サイズが小さいコーパスの代表性はさほど高くないことが主張されている。このため、本研究では図書館サブコーパスの設計方針の検討と語彙表の観察を行った。BCCWJ のマニュアル等では、語彙の偏りを防ぐためにサンプルを短くしたとの記述が見られる。そこで、サンプルを短くすれば本当に語彙の偏りが防げるのかどうかを検証するため、語彙表を使用して固定長と可変長の頻度を比較した。この結果、高頻度語はデータ量に正比例して頻度が増加するが、低頻度語は頻度がばらついて増加することが分かった。代表性が高ければ基本的にデータ量に正比例して頻度が増加するはずである。この頻度がばらつくということは、サンプル長が短い固定長の頻度が、母集団の正確な縮尺になっていないからだと考えられる。

また、低頻度語が特にばらつく理由は、固有名詞や特定のテキストに出現しやすい普通名詞が多く含まれるためだと考えられた。そこで「トマト」という普通名詞を例にサンプリング状況を観察した。「トマト」の場合、固定長では抽出範囲が短すぎ、テキスト全体における出現確率を十分に反映したサンプリングが行えていないと思われた。固有名詞や普通名詞ではこのようなサンプリングがしばしば生じていると考えられるため、高頻度語であっても一部の固有名詞や普通名詞の代表性は、それほど高くない可能性も考えられる。

ここで行った分析をさらに深める方法としては、可変長データと両方データの比較が考えられる。さらに新しい分析法としてサンプル頻度の利用も有望と思われる。現在の語彙表にはこれらのデータが不足しているため、語彙表のさらなる充実を望みたい。

文　献

Burnard, Lou（ed.）（2007）*Users' reference guide to the British National Corpus*. Oxford: Oxford University Computing Services.（http://www.natcorp.ox.ac.uk/docs/URG/を閲覧。2015.06.25）

ダグラス・バイバー、スーザン・コンラッド、ランディ・レッペン；齊藤俊雄、朝尾幸次郎、山崎俊次ほか共訳（2003）『コーパス言語学　―言語構造と用法の研究―』南雲堂.

ガイ・アシュトン、ルー・バーナード；北村裕（監訳）（2004）『The BNC Handbook　コーパス言語学への誘い』松柏社

国立国語研究所（2011）『『現代日本語書き言葉均衡コーパス』利用の手引き第 1.0 版』　国立国語研究所コーパス開発センター．http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html

丸山岳彦、柏野和佳子（2014）「サンプリング」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.21-44.

Sinclair，J. McH（1991）*Corpus, concordance, collocation*. Oxford: Oxford University Press.

田野村忠温（2014）「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.