# BCCWJ 書籍サブコーパスにおける全データ混合分析の検討

森　秀明

東北大学文学研究科

hideaki.mori.p1@dc.tohoku.ac.jp

## 1　はじめに

『現代日本語書き言葉均衡コーパス』(以下 BCCWJ) は，日本で唯一の大規模な均衡コーパスであり，高い代表性を持っていると考えられる．しかし，複雑な内部構造を持つため，真に均衡的と言えるデータは少ない [1] ．このため，真に均衡的と言えるデータのみを使用すると，データの多くを占める低頻度語で正確な頻度情報が得られない可能性がある [2] ．

BCCWJ の設計方針 [3] に反する分析法ではあるが，書籍サブコーパス（以下書籍 SC）を混合すれば頻度 11 以上の単語は 28.6%から 40.2%に，「全データ」（詳細は後述）を使用すれば 62.5%になる．そこで書籍 SC データの混合や全データの使用を行っても，正確な分析ができるかどうかを検討した．

## 2　BCCWJ の構造とデータ量

BCCWJ は総語数約 1 億語のコーパスだが，その内部は図書館 SC，出版 SC，特定目的 SC の 3 つに分かれている．図書館 SC は東京都の公立図書館の書籍を現実母集団とし，書き言葉の流通実態を捉えるために設計された SC で，語数は 3,000 万語である（数字は概数，以下同じ）．出版 SC は書き言葉を生産する書き手の立場を重視して設計された SC で，国会図書館を現実母集団にした出版書籍 SC（2,855 万語），雑誌 SC（444 万語），新聞 SC（137 万語）の 3 つからなる．特定目的 SC は上記 2 つの SC を補完する目的で，白書，教科書，yahoo! 知恵袋など 9 つの分野のデータからなる 3,500 万語の SC である．図書館 SC と出版書籍 SC はどちらも書籍をコーパス化しているため，書籍 SC とも呼ばれる．これらの SC はそれぞれに設計方針が異なり，SC ごとの語彙比較でも，SC によって出現特徴が異なる [1] [4] ．よってこれらの SC を「足し合わせて使えばよいというものではない」と指摘されている [1.p. 131.] ．

統計的な配慮を以て設計された図書館 SC，出版 SC，白書 SC はデータの収集の仕方によってさらに 3 つに区分される．字数を 1 千字にそろえた固定長と，意味のまとまりを優先して数十字〜1 万字以内のデータを抽出した可変長，およびこの両者を加えて重複を除いたデータ（これをここでは「全データ」と呼ぶ）の 3 つである．この中で統計分析に適するのは字数を一定にそろえた固定長であり，可変長はサンプルの文書によって長さが様々に異なるため，文章の論理構造の分析などに適しているとされる [3] ．出版雑誌 SC，出版新聞 SC，白書 SC の固定長は，いずれも 100 万語程度に留まるため，これらの中で一定のデータ量を有し，統計分析に適した真に均衡的な SC は，668 万語の図書館 SC 固定長と 633 万語出版書籍 SC 固定長という，二つの書籍 SC の固定長データということになる．

一方，データ量が少ないコーパスでは、正確な頻度分析が難しいことが知られている．世界で初めて作られた Brown コーパスやそのイギリス英語版である LOB コーパスの語数は 100 万語であり，この程度のデータ量では特に低頻度語で正確な頻度分析は難しいことが指摘されている [5] [6] ．その反省を元に作られ，現在のコーパスの中では高い代表性を持つと言われている British National Corps（以下 BNC）の書籍データは 4,400 万語である [7] ．コーパスの正確さはデータ量だけで決まるものではないが，BNC と比較した場合，書籍 SC 固定長のデータ量はかなり少ないと言えるだろう．

表 1 は国立国語研究所の web 上で公開されている「短単位語彙表データ」[8] に基づいて，書籍 SC 単語数を頻度ランクとデータ種別で整理した表である．

表 1　BCCWJ の書籍 SC・データ種類別単語数

| 頻度ランク | 固定長 図書館 | 固定長 出版書籍 | 固定長 重複混合 | 全データ 図書館 | 全データ 出版書籍 | 全データ 重複混合 |
|---|---|---|---|---|---|---|
| 101以上 | 4666 | 4681 | 7958 | 14535 | 13814 | 22187 |
| 51～100 | 3358 | 3167 | 5074 | 8544 | 7877 | 10453 |
| 21～50 | 7510 | 6955 | 10620 | 15733 | 14432 | 13860 |
| 11～20 | 8929 | 8349 | 10736 | 14140 | 13302 | 7036 |
| 6～10 | 11415 | 10572 | 10581 | 15765 | 15183 | 3629 |
| 2～5 | 27279 | 26180 | 13968 | 35947 | 34139 | 1772 |
| 1 | 22460 | 22489 | — | 24645 | 27025 | — |
| 単語合計 | 85617 | 82393 | 58937 | 129309 | 125772 | 58937 |
| 頻度合計(千) | 6685 | 6363 | 12927 | 30307 | 28450 | 58105 |

注：重複混合は，図書館 SC と出版書籍 SC で重複する単語の頻度を合計したもの．頻度合計の単位は千語．

頻度ランクのどこからを低頻度語と考えるかは難しいが，仮に 10 以下を低頻度語とした場合，図書館 SC 固定長では約 7 割が低頻度語となる．全データの使用や SC の混合によるメリットは大きいと言えるだろう．

## 3　単独データと混合データの比較

図書館 SC と出版書籍 SC は異なった設計方針で作られたコーパスだが，ともに書籍からデータを取得している点やサンプル数，総語数などもかなり似通っている．表 1 の頻度ランク別の単語数や単語総数を見ても，分布がよく似ているのが分かる．図書館 SC の固定長で重複している単語は約 7 割だが，頻度合計では 99.07%が重複している．重複がない約 3 割の単語はほとんどが頻度 1 の単語であり，臨時語や希少語が多い．

逆に互いのデータ量は類似してはいるものの，資料的性格としてほとんど類似点がない雑誌 SC と白書 SC の場合，重複している単語数は 26.4%，頻度合計でも 92.89%に留まる．これらに比べると図書館 SC と出版書籍 SC はかなり似通っており，「現代日本語の書籍」というより大きな母集団から抽出した 2 つの標本と見なすことも可能だろう．

同じ母集団から標本を複数抽出した場合，単独の標本の値より標本平均の方が真値に近くなる．書籍 SC が同一の母集団から抽出された標本なら，これらの平均の方が正確になるはずである．これまでのところ，コーパスの正確さを実証的に測定する方法は分かっていないが，正確なコーパスであれば全データ量の増加に対して頻度も正比例して増加すると考えられる [2]．BCCWJ では，固定長と全データという異なった分量のデータが存在している．書籍 SC の固定長と全データのデータ量の倍率はおよそ 4 倍である．これらのデータ量の倍率に対し，単語の頻度も正比例して増加しているほど，そのコーパスは正確だと考えられる．

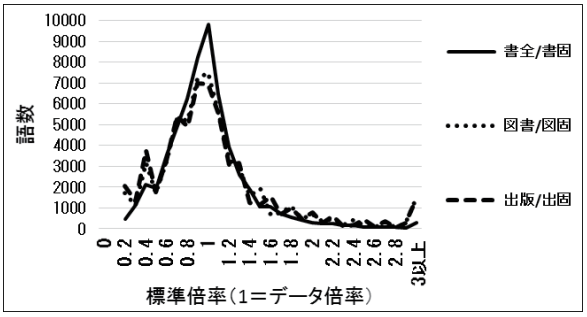


図 1 書籍単独データと混合データの比較

図 1 は図書館 SC と出版書籍 SC で重複する 58937 語について，それぞれの単語の全データ頻度÷固定長頻度がデータ量の何倍になっているかを調べ，その倍率ごとの単語数をプロットした図である．書全とは書籍 SC の重複混合全データ，書固は重複混合の固定長を意味している． 図 1 では各 SC 単独の場合より，重複混合の方が正比例する単語が多くなるのが分かる．書全/書固の平均は 0.998，標準偏差は 0.430 となっている．

表 2 データ倍率 1±0.2 の範囲で増減する単語の割合

| 頻度ランク | 重複数 | 増加数 | 減少数 | 重複割合 | 増加割合 | 減少割合 |
|---|---|---|---|---|---|---|
| 101以上 | 4601 | 81 | 11 | 98.6% | 1.7% | 0.2% |
| 51～100 | 2835 | 338 | 67 | 84.4% | 10.1% | 2.0% |
| 21～50 | 4978 | 1204 | 365 | 66.6% | 16.1% | 4.9% |
| 11～20 | 4000 | 1702 | 698 | 47.3% | 20.1% | 8.3% |
| 6～10 | 3259 | 2122 | 991 | 33.8% | 22.0% | 10.3% |
| 2～5 | 3381 | 3531 | 2068 | 20.2% | 21.0% | 12.3% |
| 1 | 615 | 2098 | 913 | 7.2% | 24.5% | 10.7% |
| 単語合計 | 23669 | 11076 | 5113 | 40.2% | 18.8% | 8.7% |

表 2 は図 1 の図書/図固と書全/書固を比較し，平均から標準偏差1倍以内に入る倍率0.8～1.2倍の単語で，どちらにも出現する単語（重複），図書/図固ではこの範囲になかった単語が書全/書固では新たにこの範囲に入った単語（増加），これとは逆の（減少）の単語数を調べ，表 1 の固定長図書館単語数の何割になっているかを求めたものである．重複割合と減少割合を合計

したものが, 図書館 SC の全データ÷固定長でもともと 1 倍前後になる単語の割合である．頻度ランクが下がるほど 1 倍前後になる割合が低くなり, 頻度 10 以下では 5 割を切る．これは頻度ランクが下がるほど誤差が大きくなるからだと考えられる．その誤差を減少させる目的で，混合を行うと，増加－減少の差の分だけ 1 倍前後になる単語が増える. 雑誌 SC と白書 SC の場合，この差は全く増加しないため, 書籍 SC は混合によって真値に近くなったと考えるのが妥当だろう（注 1）.

## 4　固定長と全データの比較

書籍 SC が同じ母集団から抽出された 2 つの標本なら，互いの単語の頻度も近い値を取るはずである．そして固定長より全データの方が正確なら，全データの類似度の方が高いはずである．[9] （pp.99-101. ）では，ウエブの検索エンジンで検索した単語頻度の分布と BCCWJ の単語頻度の分布類似度を検証するため，その対数頻度の相関を調査する方法が紹介されている．ここでは，その分析法にならい，書籍 SC で重複する各単語について，対数頻度の相関を調査する．またそれぞれの SC の頻度ランクの相関も調査する．

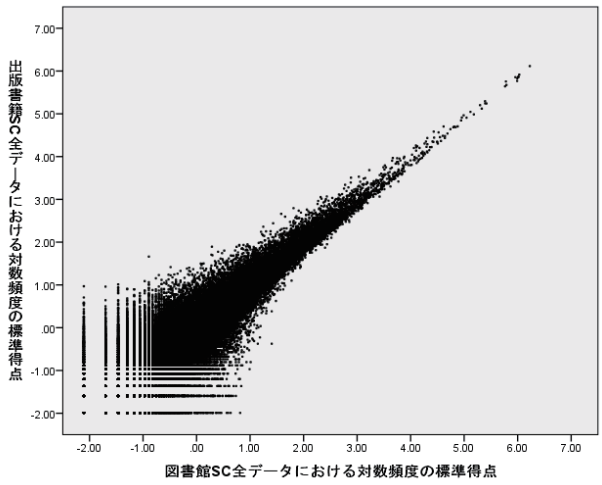


図 2　図書館 SC と出版書籍 SC 全データにおける対数頻度標準得点の散布図

初めに図 2 の散布図によって，データの分布を確認する．縦軸が出版書籍 SC，横軸が図書館 SC である．表 2 では，高頻度語ほど直線に近く，低頻度語になるほどばらつきが大きくなるのが分かる．これは低頻度語になるほど出現する文書数が少なくなり，各文書で観察される標本誤差が大きくなるためだと考えられる．

分析に当たっては，重複する各単語について，表 1 の頻度ランク別に図書館 SC 固定長の単語を区分し, 単語ごとの対数頻度については Pearson の相関係数を，頻度ランクについては Spearman のローを求めた.

表 3　図書館 SC と出版書籍 SC のデータ別相関比較

| 頻度ランク | 単語数 | Pearson の相関係数 | | Spearmanのロー | |
|---|---|---|---|---|---|
| | | 固定長の対数頻度 | 全データの対数頻度 | 固定長の頻度ランク | 全データの頻度ランク |
| 101以上 | 4666 | .942** | .945** | .896** | .900** |
| 51～100 | 3358 | .387** | .472** | .406** | .483** |
| 21～50 | 7469 | .412** | .537** | .442** | .560** |
| 11～20 | 8451 | .223** | .453** | .241** | .485** |
| 6～10 | 9642 | .149** | .458** | .153** | .473** |
| 2～5 | 16779 | .186** | .457** | .187** | .474** |
| 1 | 8572 | n.a. | .357** | n.a. | .363** |
| 単語合計 | 58937 | .858** | .883** | .788** | .856** |

**. 無相関検定の結果，1% 水準で有意 （両側）

単語合計で見ると，いずれも強い相関が観察されるが，頻度ランク別では全データの相関の方が強いことがわかる．この結果は固定長より全データの方が類似度が高いことを示唆している．全データは文字数にばらつきがあるため，統計分析に不向きだと言われる [3] ．しかし，機能語など一部の高頻度語を除けば，文字数に連動して頻度が増加する単語は多くない．内容語の多くは文書全体にまんべんなく出現することは少なく，文書の一部にあるトピックが現れるとそこから連続して使用される単語や，文字数の多い文書でも数回しか出現しない単語も多い．それらの出現数は文字数よりトピックの意味的なまとまりに左右される．可変長は意味的なまとまりによってデータを抽出しているため，これを多く含む全データの場合，文字数のばらつきによる誤差より，観察できる文書が増えたことによる正確性が上回ると考えられる．

## 5　ケーススタディ

ここまでの議論を，「長所・短所・欠点・美点」という単語を使ってケーススタディする．これらの単語は人間の性質や物の性能を評価する際によく使用される．しかし，「美点」の場合，現代では使用数が少なく，意

味も限定されていると思われる．そこでこれらの使用状況を確認するため，BCCWJ の書籍 SC で頻度を調査し，この頻度を元に全データ÷固定長の値を求めた（表 4）．図 3 はこの倍率をグラフ化したものである．

表 4 「長所・短所・欠点・美点」の頻度と倍率

| | 図書館固定長 | 出版固定長 | 図書館全データ | 出版全データ | 図全/図固 | 出全/出固 | 書全/書固 |
|---|---|---|---|---|---|---|---|
| 長所 | 49 | 60 | 225 | 264 | 4.59 | 4.40 | 4.49 |
| 短所 | 24 | 28 | 115 | 124 | 4.79 | 4.43 | 4.60 |
| 欠点 | 95 | 88 | 434 | 440 | 4.57 | 5.00 | 4.78 |
| 美点 | 5 | 8 | 39 | 23 | 7.80 | 2.88 | 4.77 |

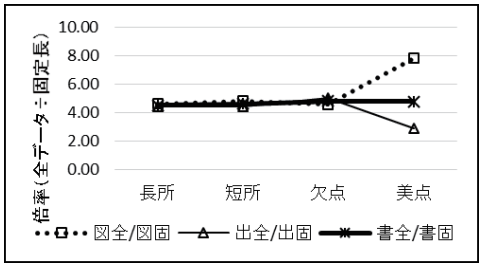


図 3 「長所・短所・欠点・美点」の倍率

表 4 の固定長における「美点」の頻度は著しく低い．図 3 で確認すると，他の 3 語が概ねデータ倍率に近い 4 倍程度になっているのに，「美点」だけは図書館 SC と出版書籍 SC で大きく値が分かれている．しかし全データを混合させると，他の 3 語とほぼ同じ倍率になる．第 4 節より，低頻度語では固定長より全データの方が正確になると思われるため，「美点」の頻度を推定するなら，書籍 SC の混合全データを使用するのが適していると考えられる．

## 6 まとめ

BCCWJ では，これまで書籍 SC の混合や全データの使用による統計分析は否定されてきた．確かに高頻度語であればサンプルの文字数を一定にした固定長頻度の信頼性が高いため，あえて全データや混合データを使用する意義は薄い．しかし，出現するサンプルが少なく，標本誤差が大きくなりやすい低頻度語の場合，サンプルの語数にばらつきがあっても，意味的なまとまりを持ったサンプルからより多くの頻度情報が得られる全データの方が，正確な値を示す可能性がある．また低頻度語の場合，類似性が極めて高い 2 つの書籍 SC を混合させた方が，より正確な値が得られる可能性がある．本論で行った分析によって，書籍 SC の全データ混合分析の正確性がすべて実証できたわけではないが，この分析法の可能性は示せたものと思う．

注 1　減少は混合させた出版書籍 SC データの誤差が大きかったため，1 倍前後の範囲から外れた可能性がある．減少の単語は個別精査し，混合の適否を判断する必要がある．

## 参考文献


[1] 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—」『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.

[2] 森秀明（2015）「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』国立国語研究所，pp.19-28.

[3] 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.

[4] 田中牧郎（2011）「語彙レベルに基づく重要語彙リストの作成 —国語政策・国語教育での活用のために— 」特定領域研究「日本語コーパス」言語政策班報告書『言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用』pp.77-88.

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

[6] 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.

[7] 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．2015.12.13）

[8] 短単位語彙表データ
http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html よりデータ取得．（2013.02.19 版）

[9] 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.