もし小学生が『現代日本語書き言葉均衡コーパス』並みに漢字を使ったら

著者	今田 水穂
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	20-29
発行年	2017
URL	http://doi.org/10.15084/00001454

もし小学生が『現代日本語書き言葉均衡コーパス』並みに漢字を使ったら

今田 水穂 (文部科学省)

What If Elementary School Students Use the Chinese Characters As Much As BCCWJ?

Mizuho Imada (Ministry of Education, Culture, Sports, Science and Technology)

要旨

『児童・生徒作文コーパス』と『現代日本語書き言葉均衡コーパス』(BCCWJ) を用いて、児童が BCCWJ と同等の水準で漢字を使用した場合に、各漢字の頻度がどの程度になるかを推定し、その結果をワードクラウドを用いて可視化した。また、その結果を用いて、学年ごとの推定頻度の比較、BCCWJ における漢字頻度との比較、教科書コーパスについて同様に漢字頻度を推定したものとの比較を行い、推定頻度と学年の相関、児童作文に固有の高頻度漢字、小学校配当外の高頻度漢字、小学校配当の低頻度漢字を調べた。

1. はじめに

児童の使用する語彙は、大人の使用する語彙とは異なる。そこで、児童の書いた作文を調査することで、児童の書き言葉の産出における漢字の需要を評価することを考える。しかし児童は基本的に学習済みの漢字しか使わず、特に低学年の場合はほとんど仮名だけで作文を書くので、単に語彙を調べただけでは、潜在的な漢字の需要を評価することができない。そこで、それらの語が大人と同等の頻度で漢字書きされた場合に、そこに含まれる漢字がどの程度の頻度になるかを試算する。

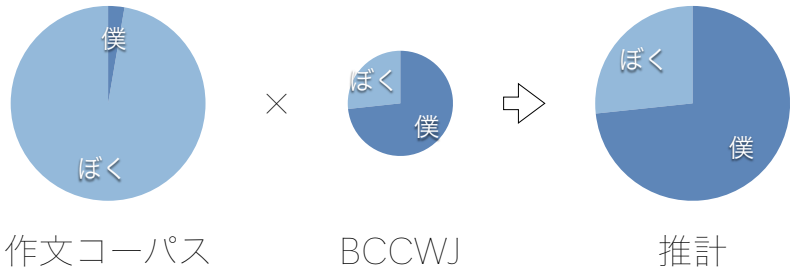

図 1　もし小学生が BCCWJ 並みに漢字を使ったら

この推計結果を用いて、児童の言語産出における漢字需要を可視化し (3 節)、学齢による漢字需要の推移 (4.1 節)、児童作文に固有の高頻度漢字 (4.2 節)、小学校配当外の高頻度漢字および配当内の低頻度漢字 (4.3 節) について考察する。

2. 方法

『児童・生徒作文コーパス』[1](作文コーパス) と、『現代日本語書き言葉均衡コーパス』[2](BCCWJ) の 2 つのコーパスを使用する。作文コーパスは小学校・中学校の児童・生徒に特定のテーマの作文課題を課し、その作文を収集・電子化したコーパスである。本調査では、2014 年度に実施した「夢」「頑張ったこと」の 2 つの作文課題について、それぞれ小学校 1〜6 年生の各 2 クラス、延べ 24 クラス分の作文資料に対して人手修正済みの形態論情報を付与したデータ[3]を使用した。

表 1　作文コーパスのサンプル数と短単位数

学年	夢		頑張ったこと		合計	
	サンプル数	短単位数	サンプル数	短単位数	サンプル数	短単位数
1	69	7196	69	10745	138	17941
2	65	11045	68	14108	133	25153
3	69	17741	69	18635	138	36376
4	78	26038	79	27481	157	53519
5	77	25265	77	29924	154	55189
6	78	26779	78	26200	156	52979
合計	436	114064	440	127093	876	241157

BCCWJ は国立国語研究所が開発した 1 億語規模の書き言葉コーパスで、13 レジスタ約 17 万サンプルの書き言葉資料によって構成される。このうち 6 レジスタ 1980 サンプル約 9 万短単位のデータがコアデータとして設定されており、この範囲のデータ全体について形態論情報の人手修正が施されている。本研究では BCCWJ コアデータのうち、他のレジスタと比べて漢字使用頻度が高い[4]新聞・白書を除いた 4 レジスタ (書籍・雑誌・ブログ・知恵袋) のデータを使用した。以下、単にコアデータというときは、この 4 レジスタを指す。

表 2　BCCWJ コアデータのサンプル数と短単位数

レジスタ	サンプル数	短単位数
書籍	83	234794
雑誌	86	241179
ブログ	471	117888
知恵袋	938	110645
合計	1578	704506

[1] 宮城・今田 (2015a)
[2] Maekawa et al. (2014)
[3] 今田水穂 (2017)
[4] 宮城・今田 (2015b)

調査は以下の手順で行った。まず、BCCWJ を使用して語別の漢字頻度表を作成した。次に作文データを使用して児童の学年別の語彙頻度表を作成した。この 2 つの数値を掛け合わせることによって、児童が BCCWJ 並みの頻度で漢字を使用した場合の漢字頻度表を作成した。この数値を以下では推定漢字頻度と呼ぶことにする。文書 a の漢字使用頻度が文書 b 並みになった時の文字 c の推定頻度を $e_{c,a,b}$ とすると、$e_{c,a,b}$ は次の式で計算できる。

$$e_{c,a,b} = \sum_{w} \frac{f_{w,a} \times g_{c,w,b}}{f_{w,b}}$$

$f_{x,y}$ は文書 y における語 x の頻度、$g_{x,y,z}$ は文書 z、語 y における文字 x の頻度である。文書 a の漢字使用頻度が文書 b 並みになった時の文字 c の 100 万字あたりの推定頻度を $\mathrm{ppm}(e_{c,a,b})$ とすると、次の式で計算できる。

$$\mathrm{ppm}(e_{c,a,b}) = \frac{10^6 \times e_{c,a,b}}{\sum_x e_{x,a,b}}$$

3. 結果

学年別の 100 万字あたり推定漢字頻度を、漢字の配当学年ごとに集計した結果を以下に示す。

表 3　100 万字あたりの推定漢字頻度

						学年
漢字分類	1 年生	2 年生	3 年生	4 年生	5 年生	6 年生
1 年配当漢字	47640	47322	47611	44966	46965	47311
2 年配当漢字	54573	57076	55670	58573	62459	63279
3 年配当漢字	38851	38364	38959	39734	42949	44269
4 年配当漢字	20867	17879	19931	22090	21998	23609
5 年配当漢字	18623	11463	11746	12535	14474	16431
6 年配当漢字	13262	13635	12116	12098	10397	13102
配当外常用漢字	22749	18668	19216	18355	18406	18103
常用外漢字	2363	1874	1687	1530	1417	1189
合計	218926	206281	206934	209882	219065	227293

全体としては 100 万字あたり 20〜23 万字が漢字であり、1 年生は例外的に漢字頻度が高いが、2〜6 年生については学年が上がるにつれて漸進的に漢字の頻度が上がることが確認できる。漢字頻度を BCCWJ 並みに調整してもこのような学年差が見られるのは、品詞や語種など語彙構成の変化を反映しているものと考える。なお、BCCWJ コアデータの漢字頻度は 100 万字あたり約 27 万字である。

個別の漢字の頻度を、ワードクラウドによって可視化したグラフを図 2 に示す。学年は低学年、中学年、高学年の 3 段階にわけ、頻度は各段階の平均を求めた。文字サイズは、頻度の平方根に比例する (従って、文字の面積と頻度が比例する)。

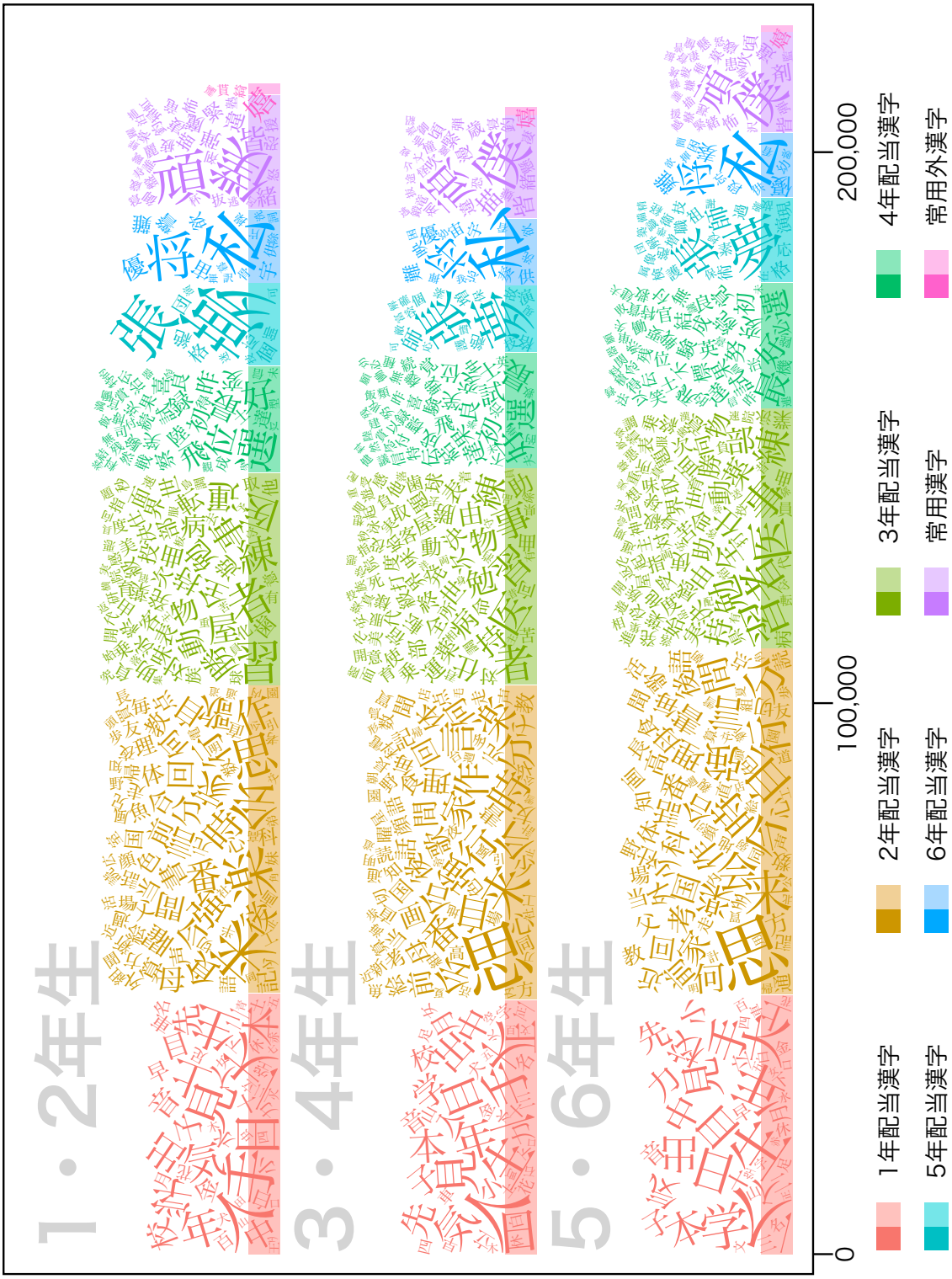
1・2年生
3・4年生
5・6年生
0
100,000
200,000
1年配当漢字
2年配当漢字
3年配当漢字
4年配当漢字
5年配当漢字
6年配当漢字
常用漢字
常用外漢字

図2　ワードクラウド

4. 考察

4.1 学年による差異

学年と推定漢字頻度の関係を調べるために、個々の漢字についてサンプルごとの推定漢字頻度を計算し、作文テーマ別に学年との相関係数を調べた。相関係数が正の値であれば学年が上がるにつれて漢字の使用頻度が上昇し、負の値であれば下降すると考えられる。図 3 は、横軸を推定漢字頻度 (全サンプル平均)、縦軸を相関係数として、各漢字を散布図で可視化したものである。頻度が 500 以上、相関係数の絶対値が 0.1 以上の漢字のみ表示する。

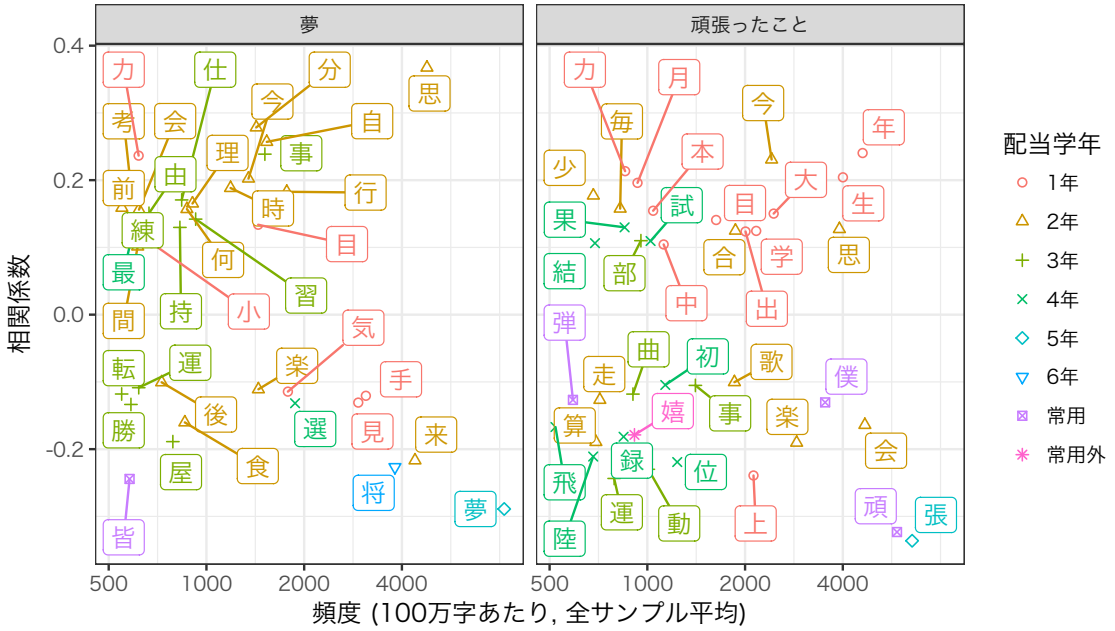

図 3　作文コーパスにおける推定漢字頻度および学年との相関係数

作文テーマによって違うが、「思、分、自、年、事、力、今、生、考」などの漢字について 0.2 以上の弱い正の相関が認められる。また、「張、頑、夢、皆、運、上、動、将、位、来」などの漢字について −0.2 以下の弱い負の相関が認められる。この結果は「思う」「考える」「自分」などの抽象的かつ一般的な語彙が学年が上がるにつれて増加するのに対して、「夢」「将来」「頑張る」「運動」など作文テーマと関連する特徴語が相対的に減少することを示唆する。減少の理由として、児童の使用語彙の変化や、1 サンプルあたりの語数の増加 (使用頻度の変化が小さい語は、相対的に単位語数あたり頻度が減少する) などが考えられる。

4.2 児童作文固有の高頻度漢字

児童作文に固有の高頻度漢字を確認するために、BCCWJ コアデータにおける 100 万語あたり漢字頻度との比較を行った。作文における推定頻度を x、コアデータにおける頻度を y とし、座標 (x, y) の原点からの距離 $\sqrt{x^2 + y^2}$ を d、x 軸からの角度 $\arctan(y/x) \times 2/\pi$ を a と

して、d を横軸、a を縦軸にプロットしたものを図 4 に示す[5]。角度が $1(=90°)$ に近いほどコアデータにおける頻度が、$0(=0°)$ に近いほど作文における頻度が高く、0.5 では同数である。

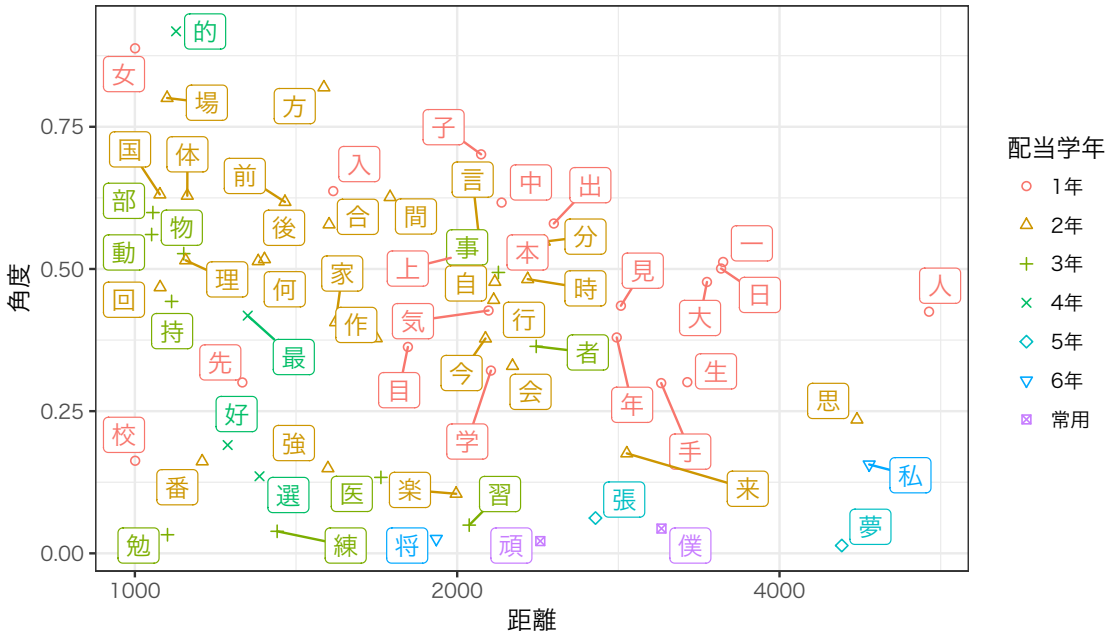

図 4　作文コーパスと BCCWJ コアデータの漢字頻度

図 4 の下方にある漢字が作文コーパスに固有の高頻度漢字と考えられるが、個々の漢字がどのような要因で児童作文において高頻度で生起するかについては、それぞれ検討を要する。「夢」「頑」「張」などは、作文テーマに固有の高頻度漢字と考えられる。「将」「来」や、医者の「医」、選手の「選」なども、作文テーマに関連した高頻度漢字である可能性がある。「私」「僕」などの 1 人称代名詞や、「思」「楽」「好」などの思考・感情語彙に含まれる漢字は、作文テーマというより生活作文などの文種に固有の高頻度漢字である可能性がある。また、「私」より「僕」の方が図の下方に位置しているのは、著者の属性 (小学生であること) の影響による可能性がある。「学」「校」「勉」「強」「練」「習」なども著者の属性に固有の高頻度漢字であろうが、このうち「勉」「強」「練」「習」などは作文テーマの影響を受けている可能性もある。これらの要因について検証するための十分な対照資料が無いため、ここでは可能性を示唆するのみに留める。

4.3　漢字の配当学年と頻度

作文は児童の言語活動の 1 つのレジスタに過ぎず、児童の漢字需要を評価するためには他のレジスタも合わせて検討する必要がある。現状、児童の言語活動を広範に調査できる均衡コーパスは存在しないが、ここでは作文コーパスの他に BCCWJ 教科書サブコーパスを使用することにする。このコーパスは BCCWJ の非コアデータに含まれるサブコーパスで、小学校か

[5] これは x を横軸、y を縦軸とする散布図について、原点を中心とする弧と両軸に囲まれた扇型の範囲を方形に変換したものに相当する。

ら高校までの検定教科書から 412 サンプル、約 93 万形態素のデータが収録されている。ここでは、小、中学校の 161 サンプル、約 36 万形態素のみを比較対象とする (以下、この範囲のコーパスを教科書コーパスと呼ぶ)。高校教科書の漢字については、中学校までに学習する漢字で対応可能であり、必ずしも小学校段階で学習する必要がないため比較対象から除外した。

小、中学校の教科書では、未履修の漢字は学習上の配慮から仮名書きに開いて表記することが多い。そのため、教科書コーパスについても作文コーパスと同様の方法で BCCWJ 並みの漢字頻度にした場合の 100 万字あたりの推定漢字頻度を計算した。作文コーパスと教科書コーパスにおける各漢字の推定頻度を用いて、小学校配当外であるが高頻度の漢字、および小学校配当であるが低頻度の漢字を調べる。

まず、高頻度の小学校配当漢字を確認する。図 5 は、作文コーパスと教科書コーパスに含まれる配当外漢字について、$d \geq 200$ のものを図 4 と同様の方法により距離と角度で表現したものである。角度が 1 に近いほど、教科書コーパスにおける頻度が高い。

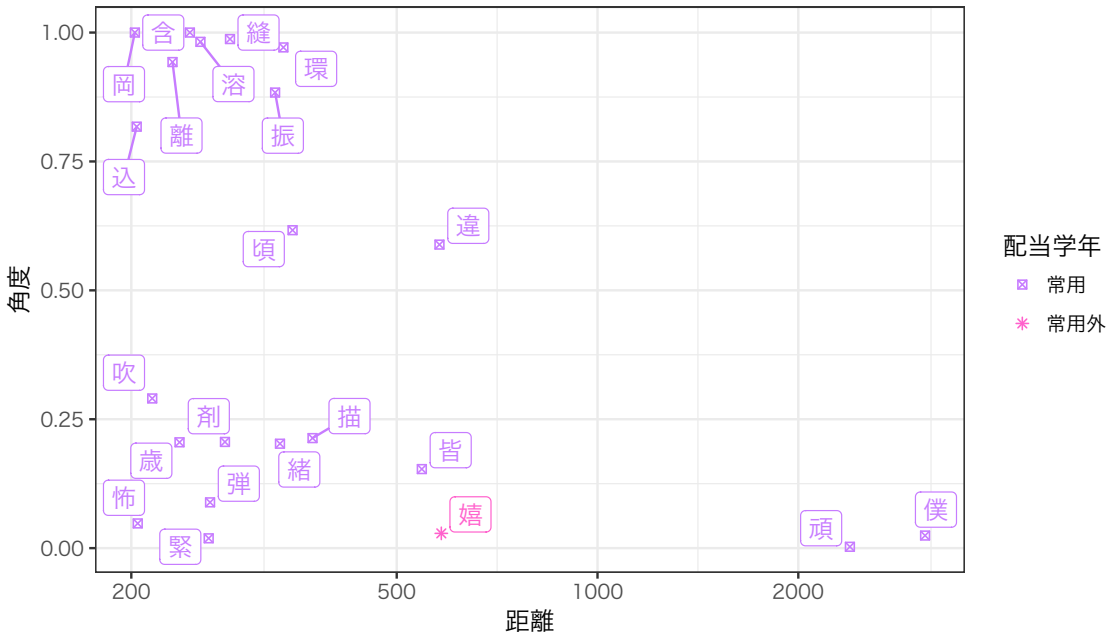

図 5　作文と教科書における高頻度の配当外漢字

$d \geq 1000$ の範囲にほとんど漢字がないことから、漢字頻度を BCCWJ 並みに調整しても、小学生の作文や小〜中学校の教科書に高頻度で生起しうる配当外漢字は少ないことが分かる。非常に頻度が高い漢字としては「僕」「頑」があるが、作文コーパスのみ高頻度で、教科書コーパスでは低頻度である。「頑」は作文テーマの影響で頻度が高くなっているものと考えるが、「僕」は本調査資料に限らず児童の書き言葉では多用される可能性があり、小学 6 年配当の「私」と合わせて学習時期を検討する余地のある漢字と言える。また、やや頻度は下がるが、作文コーパス、教科書コーパスの両方で頻度が高い「違」「頃」などについても、小学校で学習したとしても不自然ではないと考える。

次に、低頻度の小学校配当漢字を確認する。図 6 は、作文コーパスと教科書コーパスに含ま

れる小学校配当漢字について、$d < 20$ のものを表示したものである。図には含まれていないが、作文コーパス、教科書コーパスのいずれも頻度 0 だった配当漢字として、小学 5 年配当の「俵」、小学 6 年配当の「絹」「蚕」がある。

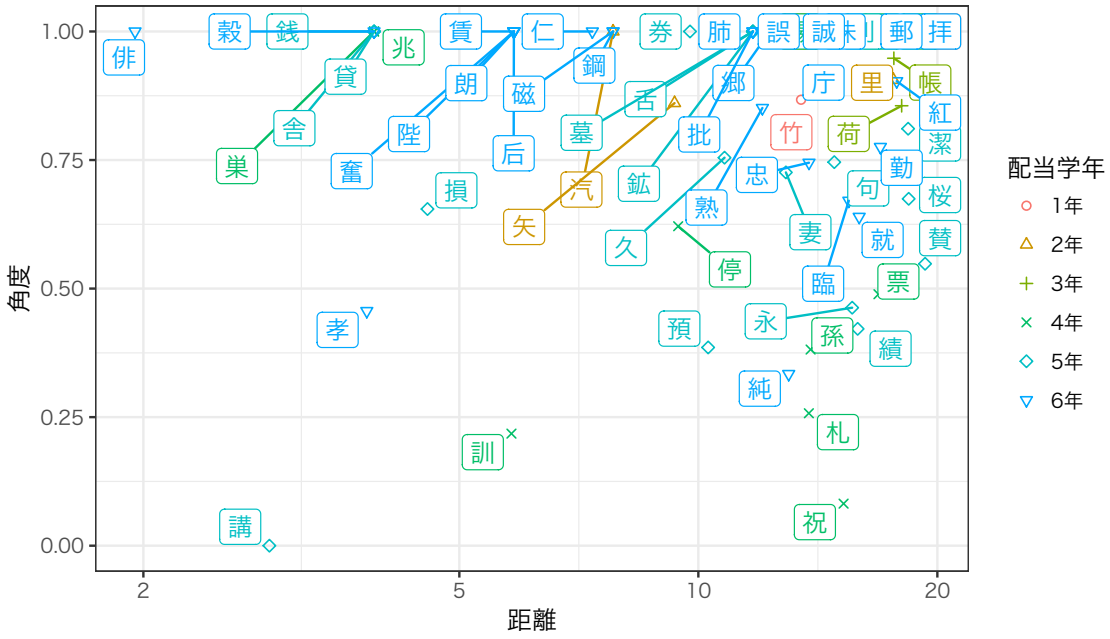

図 6　作文と教科書における低頻度の配当漢字

図中の漢字の頻度は 100 万字中 20 字以下であり、非常に頻度の低い漢字と言うことができるが、教育漢字の需要は必ずしも使用頻度のみで評価できるものではない。例えば俳句の「俳」「句」、音訓の「訓」、熟語の「熟」、批評の「批」などは国語の学習において必要になる漢字であり、頻度が低いからといって重要度が低いとは断定できない。また、小学 1 年配当の「竹」、小学 2 年配当の「矢」「里」なども頻度は低いが他の漢字の構成要素となる字であり、早い段階で教えることは一定の合理性がある。一方で、小学 2 年配当の「汽」や、前述の「俵」「絹」「蚕」などのように、必ずしもこの段階で学習する必要があるかどうか、検討の余地のある漢字も見られる。

5.　まとめ

『児童・生徒作文コーパス』と『現代日本語書き言葉均衡コーパス』の 2 つの言語資源を利用して、児童が大人と同等の使用頻度で漢字を使用した場合の推定漢字頻度を試算し、その結果の可視化と、学年差、レジスタ差、漢字の配当学年と推定頻度の関係などについて検討した。本研究で得られた知見を以下に列挙する。

- 学年差について、児童作文における推定漢字頻度は 100 万字あたり 20〜23 万字ほどで、BCCWJ コアデータにおける 27 万字よりも少なく、学年が上がるにつれて増加する傾向がある。個別の漢字を見ると、「思」「考」など学年が上がると推定頻度が増加する漢字がある一方で、「夢」「頑」「張」など作文テーマに直結する漢字は相対的に推定頻度が

低下する。

- レジスタ差について、BCCWJ と比べて児童作文に固有の高頻度漢字の中には、作文テーマ、文種、著者の属性など様々な要因の影響を受けていると考えられるものが混在している。
- 配当学年と推定頻度の関係について、「僕」「違」「頃」など配当外漢字の中にも作文や教科書において高頻度で使われうる漢字がある一方で、「汽」「俵」「絹」「蚕」など配当漢字の中にも非常に頻度の低い漢字がある。

BCCWJ を利用して教育漢字や常用漢字の分析をした研究としては、これまで棚橋 (2013)、丹保 (2014, 2016)、河内 (2015) などがある。特に丹保 (2014, 2016) は BCCWJ における高頻度漢字、低頻度漢字について配当表漢字としての妥当性を検討しており、本研究と目的、方法の重なる点が多い。

先行研究に対する本研究の新規性は、児童作文という児童の産出言語を資料として使用したこと、またその分析手法を提案したことである。資料について、児童作文は既存の他の資料にはない特徴を持つ。例えば丹保 (2016) が BCCWJ における高頻度漢字として挙げている「彼」は、本研究で使用した BCCWJ コアデータにおいても 100 万字あたり 682 字ほどで配当外漢字としては最も頻度が高いが、作文コーパスでは 20 字、教科書コーパスでは 75 字ほどと低頻度である。作文や教科書以外のレジスタも調べる必要があるが、単に大人の文章で頻出するというだけであれば小学校までに学習する必然性はなく、中学校までに学習する常用漢字に含まれていれば十分である。一方、「僕」はコアデータでは 212 字、教科書では 119 字ほどの頻度だが、作文コーパスでは 3096 字と突出して高い。大人の文章や学習教材だけを調査対象としてしまうと、このような児童の生活に固有の漢字需要を見落とす恐れがある。

また分析手法について、児童作文を対象とした漢字需要調査は、児童の漢字使用状況が既存の教育カリキュラムの影響を受ける (未履修の漢字は生起しない) という難しさがある。習得後はほぼ漢字表記されるような漢語や専門語彙であれば、語彙を調べることで漢字の需要もほぼ特定することができるが、例えば「あいつ」などの語は大人の文章でも「彼奴」と書くことは稀であり、単に全ての語彙を漢字表記に置き換えることで漢字の需要を数値化することはできない。この問題に対して、本発表は BCCWJ における漢字頻度を用いて潜在的な漢字需要を推定するという手法を提案した。この手法により、児童の漢字需要を評価するために一定の成果を示せたものと考える。

学習漢字の妥当性が頻度だけでなる様々な観点から複合的に評価すべきものであることは、先行研究の全てに共通する見解である。丹保 (2016) も、BCCWJ における頻度のみならず様々な観点から検討を行い、「彼」は高頻度漢字ではあるが用法が限られているため、配当表漢字にはふさわしくないと結論している。しかしながら、漢字の使用頻度や潜在的な需要も、その漢字の重要度を評価するための主要な指標の一つであることは疑いない。本研究で利用した作文資料は特定のテーマに沿って書かれたものであるため、児童の書き言葉の全体に対する代表性という観点からは問題の残る部分もあるが、学習漢字の評価を考える上で従来なかった新たな観点を提案するものとして、今後の研究における参考の一つとなることを期待する。

謝　辞

本研究は JSPS 科研費 JP16H00011 の助成を受けたものです。本研究で利用した言語資源のうち、『現代日本語書き言葉均衡コーパス』は国立国語研究所が開発した言語資源です。『児童・生徒作文コーパス』の本文は科研費基盤 (B)「言語研究の実践的応用に関するリサーチユニット」(代表: 矢澤真人)、形態論情報の一部は漢検研究助成「作文コーパスを資料に児童・生徒の漢字使用・選択傾向と発達の実態を明らかにする」(代表: 宮城信) による成果物です。データの利用を許諾いただいた各位に感謝します。

文　献

宮城信・今田水穂 (2015a).「『児童・生徒作文コーパス』の設計」　第 7 回コーパス日本語学ワークショップ予稿集, pp. 223–232.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). "Balanced Corpus of Contemporary Written Japanese." *Language Resources and Evaluation*, 48, pp. 345–371.

今田水穂 (2017).『『児童・生徒作文コーパス』形態論・係り受け情報データ』, (バージョン 1.3) (2017 年 2 月作成).

宮城信・今田水穂 (2015b).「『児童・生徒作文コーパス』を用いた漢字使用能力の推定」　第 8 回コーパス日本語学ワークショップ予稿集, pp. 47–56.

棚橋尚子 (2013).「学年別漢字配当表に配当された漢字と習得語彙との関係」　全国大学国語教育学会発表要旨集　125 巻, pp. 307–310.

丹保健一 (2014).「学年別漢字配当表の字種選定を巡って: 頻度下位の 10 字種を中心に」　三重大学教育学部研究紀要, 65, pp. 73–90.

丹保健一 (2016).「学年別漢字配当表の字種選定に関する基礎的研究: 使用頻度上位の非「配当表漢字」10 字種を巡って」　三重大学教育学部研究紀要, 67, pp. 33–48.

河内昭浩 (2015).「国語教育のための「常用漢字表」語例の検討」　第 7 回コーパス日本語学ワークショップ予稿集, pp. 113–122.

関連 URL

発達段階と到達目標を考慮した学齢別漢字重要度評価法の開発

https://sites.google.com/site/kaken16H00011/

作文を支援する語彙・文法的事項に関する研究プロジェクト

https://sites.google.com/site/sakubunshienpropject/

現代日本語書き言葉均衡コーパス (BCCWJ)

http://pj.ninjal.ac.jp/corpus_center/bccwj/