学校課題作文コーパスの構築

今田　水穂 (筑波大学)
宮城　信 (富山大学)

Construction of a corpus of assigned school essays

Mizuho Imada (University of Tsukuba)
Shin Miyagi (University of Toyama)

要旨

児童の作文能力を研究するための資料整備を目的として、現在の児童の作文調査や、過去の作文資料の電子化を進めている。この研究の一環として、国語研究所所蔵の 1980 年代の作文資料（島村 1987）を電子化したので、その概要を報告する。この資料は昭和 58 年に千葉県内の公立小学校 2 年、4 年、6 年の児童の作文を調査したもので、「学校」「先生」「ともだち」の 3 つの課題を含む。原資料は約 1440 編ほどの規模の調査と考えられるが、資料の欠落もあり、電子化した資料は 1021 編である。資料の概要と電子化作業の詳細について報告し、既に構築済みの「児童・生徒作文コーパス」(2014-2016)、「「手」作文コーパス」(1992, 2016) との違いについて、文字種の構成比を中心に説明する。

1. はじめに

2017 年に学習指導要領が改訂され、小学校では 2020 年度から新要領が実施された。新要領では語彙学習に関する項目が明確化され、低学年、中学年、高学年でそれぞれ「身近なことを表す語句」「様子や行動、気持ちや性格を表す語句」「思考に関わる語句」の量を増やし、「語彙を豊かにする」ことが明記されている。語彙教育の改善のために児童の語彙使用の実態を把握することは不可欠であるが、児童が生活の中で使用している語彙を直接観察することは難しく、書き言葉に限定されるが観察可能な資料として児童作文の有用性は大きい。

研究資料としてのコーパスは、十分な規模と均衡性を有していることが望ましい。特に語彙の量的な研究では、小規模なコーパスでは低頻度の語について有意な傾向を見出すことは難しく、ある程度の規模のコーパスを使用することが不可欠である。また、作文の文種やテーマは語彙の選択に大きく影響するため、多数の文種、テーマの作文が含まれていること、および、各作文の文種、テーマなどのメタデータが記録されており、均衡を保証できることが必要である。

宮城・今田 (2018) は、2014 年度から 2016 年度までの 3 年間にかけて、小学校 1 年生から中学校 3 年生までの作文を悉皆的に調査・収集した資料を電子化し、165 万形態素規模の「児童・生徒作文コーパス」を構築した。このコーパスは語彙研究のために十分な規模を有しているが、調査校が特定の国立大学附属校である点、作文テーマが「夢」「頑張ったこと」の 2 種類のみである点などから、十分な代表性を有しているとは言えない。同様の調査をより大規模に実施することは容易ではなく、過去に実施された大規模調査の資料を電子化し、作文研究のた

めのコーパスを整備することも併せて考える必要がある。阿部ほか (2017) は、1992 年に小学校 1 年生から中学校 3 年生までの作文を悉皆的に調査・収集した資料 (成田ほか 1995) に、同条件で 2016 年に調査・収集した資料を併せて電子化した 28 万形態素規模の「「手」作文コーパス」を構築した。このコーパスは「児童・生徒作文コーパス」とは別の国立大学附属校で、「手」というテーマで書かれた作文コーパスであり、「児童・生徒作文コーパス」とは別のテーマの作文を補完することができる。

これらに加えて、本研究では国立国語研究所が所蔵する島村 (1987) の調査資料を電子化し、34 万形態素規模のコーパスを構築した。この資料は 1983 年に 3 つの公立小学校の 2 年生、4 年生、6 年生を対象として調査・収集されたものであり、「学校」「先生」「友達」というテーマで書かれた作文である。調査学年数が少なく、個々のテーマの標本数は多くはないが、前述の 2 つのコーパスとは別のテーマの作文を補完する資料として価値が大きい。本発表では、このコーパスの概要を報告し、文字種の構成比を中心として既存の 2 つのコーパスとの違いを説明する。

2. コーパスの構築

国立国語研究所の所蔵していた昭和期の作文資料を電子化した。資料には詳細な記録が付属していなかったが、島村 (1987) で報告されたものと考えられる。島村 (1987) によると、この資料は 1983 年の 2 月と 3 月に、千葉県松戸市の公立小学校 3 校の 2 年生、4 年生、6 年生を対象として、「わたしの学校」(以下「学校」)「先生」「ともだち」の 3 つの題で作文を書かせたものである。ただし「ともだち」は 2、3 週間の間隔を開けて 2 回書かせており、延べ 4 回の調査を実施している。いずれの調査校も各学年 1 学級、3 学年で約 120 名の生徒を調査対象としており、4 回の調査で 480 編の作文を収集している。それを 3 つの調査校で実施しているので、当時収集した資料の規模は 1440 編程度と推測される。この資料を人手でコンピュータに入力し、テキストファイルを作成した。

3. コーパスの概要

3.1 コーパスの規模

資料には欠落があり、特に「ともだち」の 2 年生の資料は全て逸失していた。最終的に、電子化した資料は 1021 編である。課題別、学年別の作文数を表 1 に、総形態素数を表 2 に示す。括弧内は女子の内数で、形態素数は MeCab/UniDic で自動解析したものである。全体の作文数は 1021 編、形態素数は約 34 万で、どの課題、学年とも、おおむね男女比は半々である。

表 1　コーパスに収録した作文数 (課題、学年別)

theme	小 2	小 4	小 6	Total
学校	119 (63)	113 (47)	69 (35)	301 (145)
先生	114 (58)	113 (56)	96 (52)	323 (166)
友達 1	0 (0)	110 (53)	76 (37)	186 (90)
友達 2	0 (0)	109 (54)	102 (46)	211 (100)
Total	233 (121)	445 (210)	343 (170)	1021 (501)

表 2　コーパスに収録した作文の形態素数 (課題、学年別)

theme	小 2	小 4	小 6	Total
学校	21566 (10475)	33219 (13167)	22804 (9075)	77589 (32717)
先生	28828 (13250)	37187 (16523)	36743 (16731)	102758 (46504)
友達 1	0 (0)	46327 (21561)	32296 (14321)	78623 (35882)
友達 2	0 (0)	41812 (19388)	37072 (13425)	78884 (32813)
Total	50394 (23725)	158545 (70639)	128915 (53552)	337854 (147916)

3.2　作文の長さ

課題、学年、男女別の作文の長さ (文字数) を、課題、学年、男女別に図 1 に示す。このデータは、段落頭の字下げなどの空白文字を含む、ファイル内の実際の文字数である (改行文字は含まない)。島村 (1987) では、400 字詰め原稿用紙で作文を書かせたことが報告されているが、枚数制限の有無は記述されていない。データを見ると 2000 字近く書かれている作文もあり、枚数制限は無かったものと推測できる。

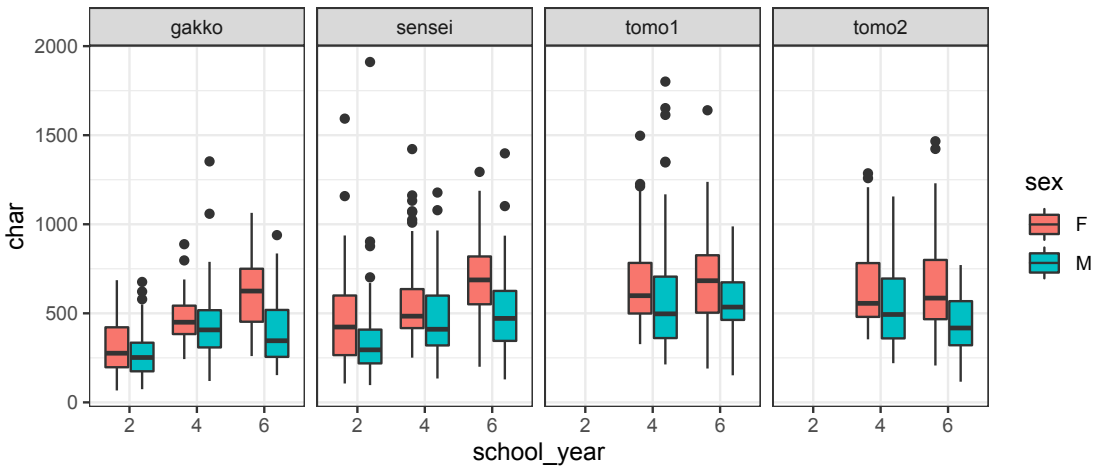

図 1　作文の文字数 (課題、学年、男女別)

4. 文字種の分析

4.1 対照コーパス

構築したコーパスの文字種について検討する。比較対象として、「児童・生徒作文コーパス」(宮城・今田 2018) と「「手」作文コーパス」(阿部ほか 2017) を使用する。前者は 2014 年から 2016 年にかけて調査した国立大学附属小学校・中学校の作文コーパスで、「夢」「頑張ったこと」という 2 つのテーマの作文を含む。後者は 1992 年と 2016 年に同一の国立大学附属小学校・中学校で調査した作文コーパスで、「手」というテーマの作文である。本発表では、これらのコーパスから小 2、小 4、小 6 のデータのみを使用し、本研究で構築した作文コーパスと比較する。これらのコーパスの課題別、学年別の作文数を表 3 に示す。括弧内は女子の内数である。

表 3　対照コーパス (課題、学年別)

survey_year	theme	小 2	小 4	小 6
1992	手	40 (20)	40 (19)	40 (20)
2016	手	101 (51)	98 (49)	126 (64)
2014-2016	頑張ったこと	203 (103)	217 (108)	228 (111)
2014-2016	夢	204 (101)	216 (107)	227 (111)

作文の長さ (文字数) については割愛するが (先行研究を参照されたい)、本研究で構築したコーパスの原資料が字数制限をされていないと推測されるのに対して、これらのコーパスの原資料は原稿用紙 1 枚という条件で書かれており、おおむね 400 字に収まる文字数である (ただし、欄外などを使って書いたと思われるものがあり、400 字を超えるものもある)。

これらのコーパス、および本コーパスは多様なテーマの均衡性のみならず経年比較も想定して整備されたものだが、作文の特徴には経年差以外の様々な属性が影響し得ることにことに注意されたい。例えば、学校の設置者 (国立、公立、私立) の違いによる入学者層の違いや、作文テーマ、調査時の条件 (時間や原稿用紙の枚数など)、性別などの属性は、語彙の選択や語数、文字数、その他の特徴に影響を及ぼす。そのため、たとえ 1983 年と 2016 年の作文に何らかの特徴差が認められたとしても必ずしも年代差が原因とは限らない。影響し得る属性を可能な限り注意深く検討する必要がある。

以下では漢字や仮名などの文字種の頻度についてコーパス間の比較を行うが、作文の長さの違いを考慮し、素頻度ではなく率で比較する。また、本発表は性別差の分析を目的としたものではなく、またどのコーパスも男女比はおおむね半々なので、簡単のために性別の区別はせず単に合算する。それ以外の属性については、データを見ながらその影響を検討する。

4.2 文字種の内訳

コーパス全体のおおまかな文字種の内訳を確認するため、作文テーマ別に文字種の内訳を集計した (図 2)。文字種は、特に漢字やひらがなについて学年による頻度の違いが大きい (高学年

ほど漢字の頻度が高い) と予想されるが、1983 年の「友達」は 2 年生のデータがないため、他の課題についても 2 年生を除外して 4 年生と 6 年生のデータを合算して比較した。文字種は、Unicode 文字プロパティの Script 名に基づいて分類した。全てのコーパスを通じて、使用されていた文字は記号類 (Common)、漢字 (Han)、ひらがな (Hiragana)、カタカナ (Katakana)、ラテン文字 (Latin) の 5 種のみだった。全体としては、ひらがなの率が 6 割を超えており、次いで漢字、記号、カタカナが多く、ラテン文字はごくわずかに確認されたのみである。1983 年の資料と他の資料を比べると、前者の方が記号の率が高い。また、1983 年の「友達」は漢字の率が低い。カタカナは、2014〜16 年の「夢」「頑張ったこと」が多く、1992 年と 2016 年の「手」が少ない。1983 年の資料は、その中間である。

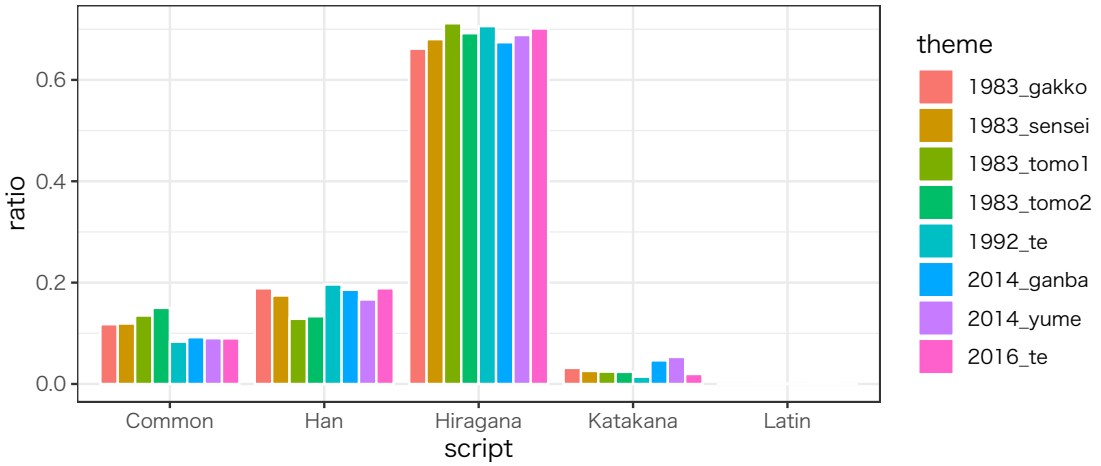

図 2　文字種 (Unicode Script) の内訳 (課題別, 4 年生 + 6 年生)

4.3　記号

1983 年の作文で記号 (Common) の使用率が高い原因を特定するために、どの記号が頻度差に影響しているかを調べた。まず、1983 年資料 (4 年と 6 年) の記号の内訳を Unicode 文字プロパティの一般カテゴリ (General Category) ごとに集計した。その結果、記号 58,088 字中、Other_Punctuation が合計 52,930 字と突出して多く、次いで括弧 (Open_Punctuation、Close_Punctuation)、修飾文字 (Modifier_Letter)、数字 (Decimal_Number) が多かった。また文字別に見ると、読点、句点、アスタリスク (＊)、長音記号 (ー) が多かった。このうち「＊」は、2 つのコーパスで個人情報などをマスクするための伏字として使用されている文字である。

表 4　記号 (Common) の内訳 (1983 年資料のみ、4 年 + 6 年)

general_category	char	freq	Total
Other_Punctuation	、	23107	52930
Other_Punctuation	。	14712	
Other_Punctuation	＊	14079	
Other_Punctuation	<other>	1032	
Decimal_Number	<other>	1104	1104
Modifier_Letter	ー	1327	1327
Open & Close_Punctuation	<other>	2429	2429
<other>	<other>	298	298
Total	-	58088	58088

これらの高頻度の文字が、課題ごとの記号使用頻度の違いに強く影響していることが考えられる。そこで、これらの文字がテキスト中に占める割合を課題ごとに調べた。結果を以下に示す。

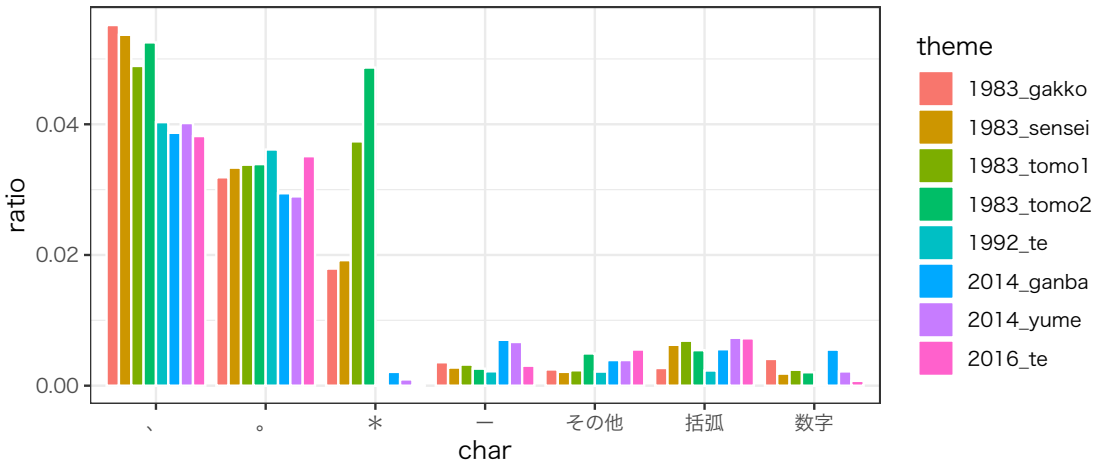

図 3　記号 (Common) の内訳 (課題別, 4 年 + 6 年)

1983 年の資料は他の資料と比べて、読点と＊の率が高い。特に 1983 年の「ともだち」では＊の率が顕著に高い。これらが 1983 年の資料に記号が多い主要因だと考えられる。読点が多い理由については、課題の影響、学校の指導方針の影響など、様々な可能性が考えられるが、現状特定することができない。＊は、これらのコーパスを構築する際に個人情報を隠すために伏字として使用した文字である。「ともだち」は作文中に友人の名前を書くことが多かったために伏字が顕著に多く、「学校」や「先生」も学校名や先生の名前を書くことが多かったために伏字が多かったと推測できる。また、伏字は通常「＊＊＊＊」のように複数文字の列として使用されるので、単独で使用される他の記号と比べて文字数の差が顕著に現れたことが考えら

れる。

それ以外の目立った特徴としては、句点は 2014〜2016 年の資料で少なく、1994 年と 2016 年の「手」作文で多い。句点の頻度の違いは、文の長さの違いを示唆する。句点が少ない「夢」「頑張ったこと」は文が長く、句点が多い「手」は文が短い。図 2 で見たように「夢」「頑張ったこと」はカタカナが多く、「手」はカタカナが少ないため、カタカナ語の頻度が文の長さに影響していることが考えられる。長音記号は「夢」「頑張ったこと」で多いが、これもカタカナの頻度が高いことと関係していると思われる。

括弧は大半がカギ括弧である。「先生」や「ともだち」で括弧が多いのは人物の発話が多いためだと思われるが、「夢」「頑張ったこと」で多い理由や、「手」の 1993 年と 2016 年で頻度が違う理由は不明である。数字は 1983 年の「学校」と 2014 年の「頑張ったこと」で多い。前者は学校生活に関する話題で学年、時間、人数などを表すために、後者はスポーツや勉強に関する話題で、時間、距離、得点、順位などを記述するために使用されているようである。

4.4 カタカナ

カタカナは、2014〜2016 年の資料で頻度が高く、1992 年と 1996 年の「手」で頻度が低い。カタカナは外来語などのカタカナ語に多く含まれると思われ、外来語の多くは名詞である。そこで、形態論情報を用いて名詞の語種の内訳を調べた (図 4)。1983 年のデータは mecab と UniDic で形態論情報を自動付与した。「手」は 1992 年、2016 年とも形態論情報を自動付与後、人手で修正済み、「夢」「頑張ったこと」は部分的に人手修正済みである。小学校低学年の作文では単語を部分的にのみ漢字書きする例が多く、自動付与による形態論情報の精度が低いため、4 年と 6 年のデータのみ合算して集計した。図中の固は固有名詞、混は混種語、記号はごく少ないが「DVD」や「CD」のような略語、「m」のような単位などである。

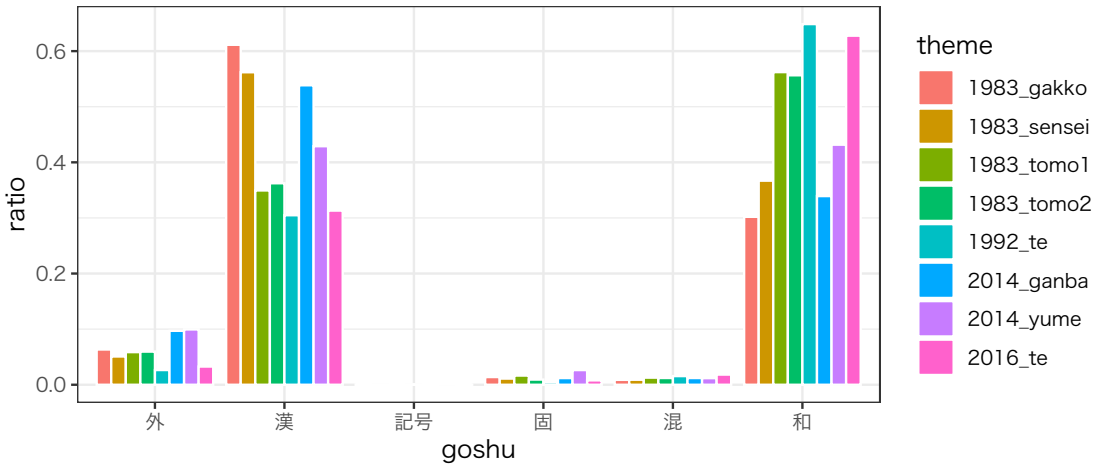

図 4　名詞の語種の内訳 (課題別、4 年 + 6 年)

予想の通り、2014〜2016 年のデータは他のデータよりも外来語の頻度が高く、1992 年と 2016 年の「手」は外来語の頻度が低い。この使用語彙の違いが、コーパス間のカタカナの使用頻度の違いの主要因だと考えられる。

それ以外では、1983 年の「友達」と 1992 年、2016 年の「手」で漢語の使用率が低く、和語の使用率が高い。これらは課題の「友達」「手」という語が和語であることが強く影響しているものと考えられる。同じく和語の「夢」も、「友達」「手」ほどではないが漢語の使用率がやや低い。図 2 を見ると、「手」は漢語の頻度の低さにも関わらず、漢字の頻度はそれほど低くない。これは、和語の「手」が漢字書きされる頻度が高いことを示唆している。

カタカナ語について、実際にどのような外来語が使われているかを確認するために、各作文テーマで使われている外来語の高頻度語彙を調べた。上位 10 位までを図 5 に示す。数値は、名詞に占める割合である。

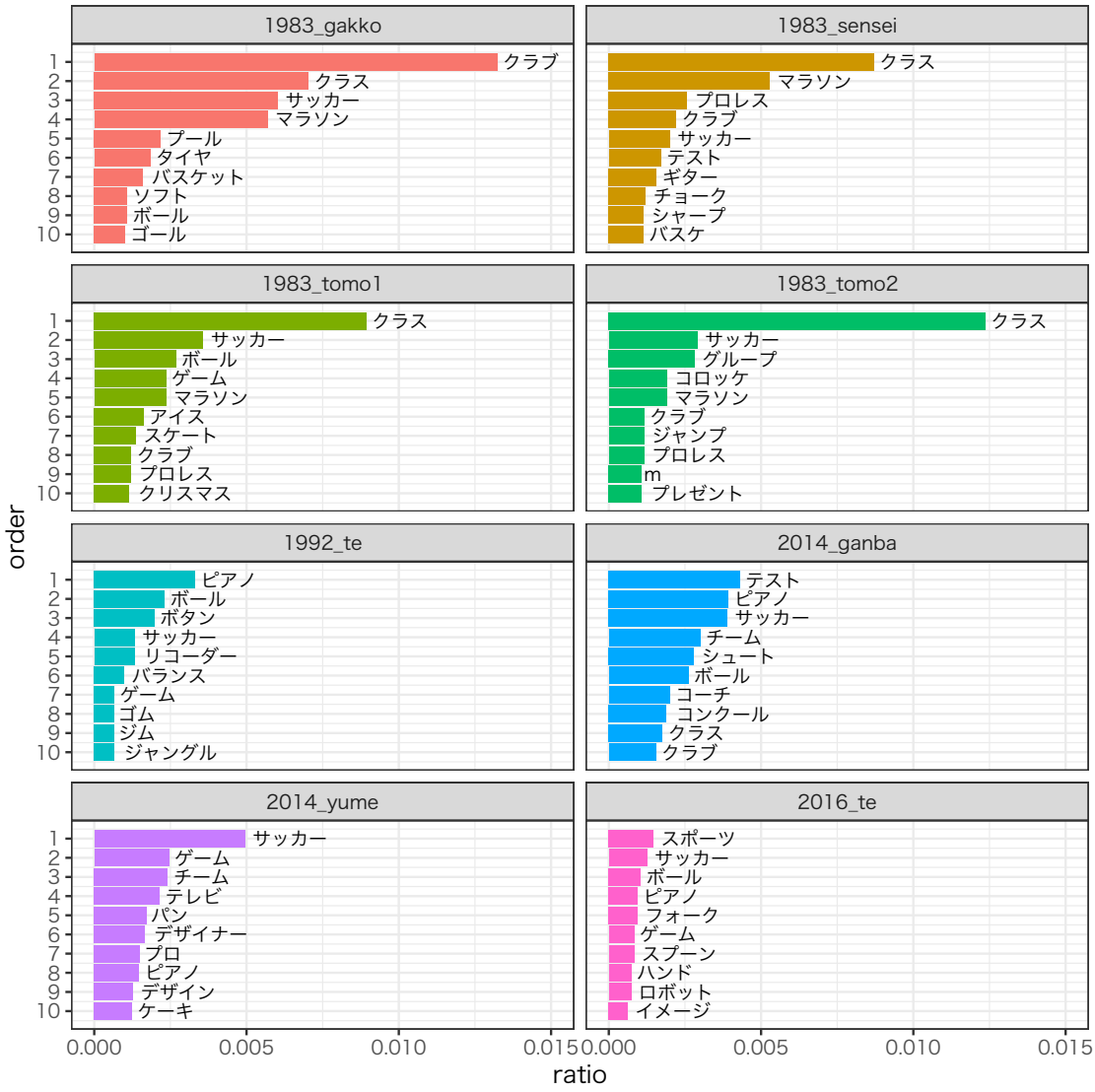

図 5　高頻度の外来語 (課題別、4 年 + 6 年)

外来語が多い 2014〜2016 年の資料では、「頑張ったこと」では勉強やスポーツ、習い事に関する語彙が、「夢」では職業に関する語彙が多く見られる。1983 年の「学校」「先生」「友達」で

は「クラス」など学校生活に関する一部の語彙が非常に多いが、全体としては 2014〜2016 年の資料より外来語は少ない。1992 年と 2016 年の「手」は手を使った活動に関する外来語が見られるが、突出して頻度が高い語も無く、全体として外来語の頻度は低い。これらのことから、カタカナの使用頻度の差は、主に課題に応じた語彙の選択の影響によるものと考えられる。

4.5 漢字

漢字は、学年が上がるほど頻度が上昇する傾向が顕著であり、また使用する漢字は学年別漢字配当表の影響を強く受ける。また、作文テーマの違いによって、使用する語彙に違いが生じ、それが漢字の使用頻度に影響を及ぼす。そこで、まずは課題別、学年別の漢字の使用頻度を確認する (図 6)。ここでは漢字とかなの使用比率を見たいので、記号とラテン文字を除外し、漢字、ひらがな、カタカナの和を分母とする割合で漢字の頻度を評価する。また、学年別配当漢字の内訳も併せて確認するため、これらを色分けして積み上げ棒グラフで可視化する。年代によって常用漢字も学年別配当漢字も異なっているので、それぞれの時代の区分で分類した。1983 年資料は常用漢字表 (1981 年) と漢字配当表 (1977 年)、1992 年資料は常用漢字表 (1981 年) と漢字配当表 (1989 年)、2014 年以降の資料は常用漢字表 (2010 年) と漢字配当表 (1989 年) を使用している。図中の 1〜6 が学年別配当漢字、8 がそれ以外の常用漢字、9 がその他の漢字である。

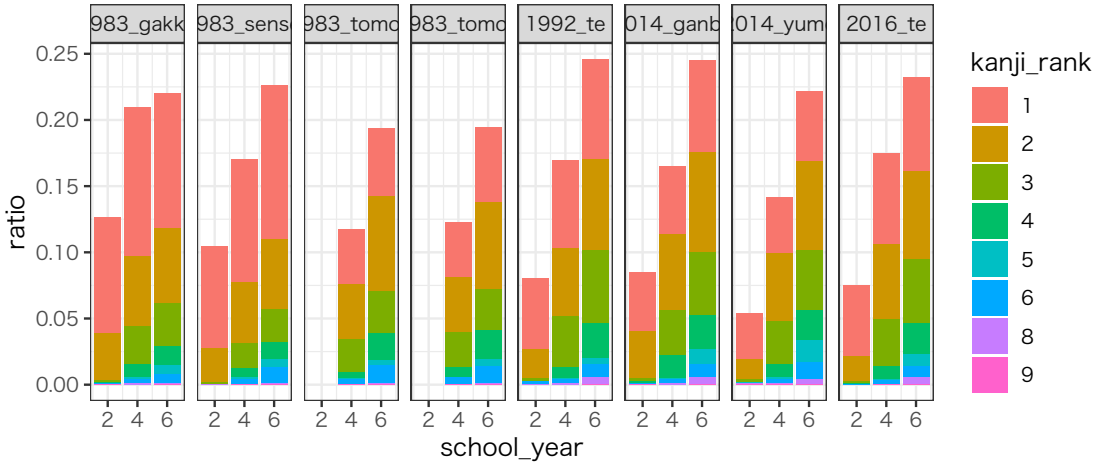

図 6　漢字の内訳 (課題別・学年別)

全体として、学年が上がるほど漢字の頻度は高くなっている。また、2 年生ではほとんど 1〜2 年次配当漢字しか使われていないが、4 年生では 3〜4 年次配当漢字が使われ始め、6 年次ではさらに 5〜6 年次配当漢字が使われるようになるなど、配当漢字が児童の漢字使用に影響を及ぼしていることが分かる。一方で、6 年生においても半数以上の漢字は低学年配当漢字であり、配当学年の低い漢字ほど頻度が高い。「ともだち」は他の課題と比べて漢字の使用率が低いが、前述の通りこの課題は他の課題より伏字が多く、伏字の中に固有名などの漢字が多く含まれていたことが考えられる。「夢」「頑張ったこと」はカタカナが多く、「手」はカタカナが少ないが、その影響は漢字の使用率に顕著な影響は及ぼしていないように見える。

次に、各学年別配当漢字の頻度がどの学年で増えるかを確認するため、6 年次の漢字使用率を 1 とする折れ線グラフを示す (図 7)。

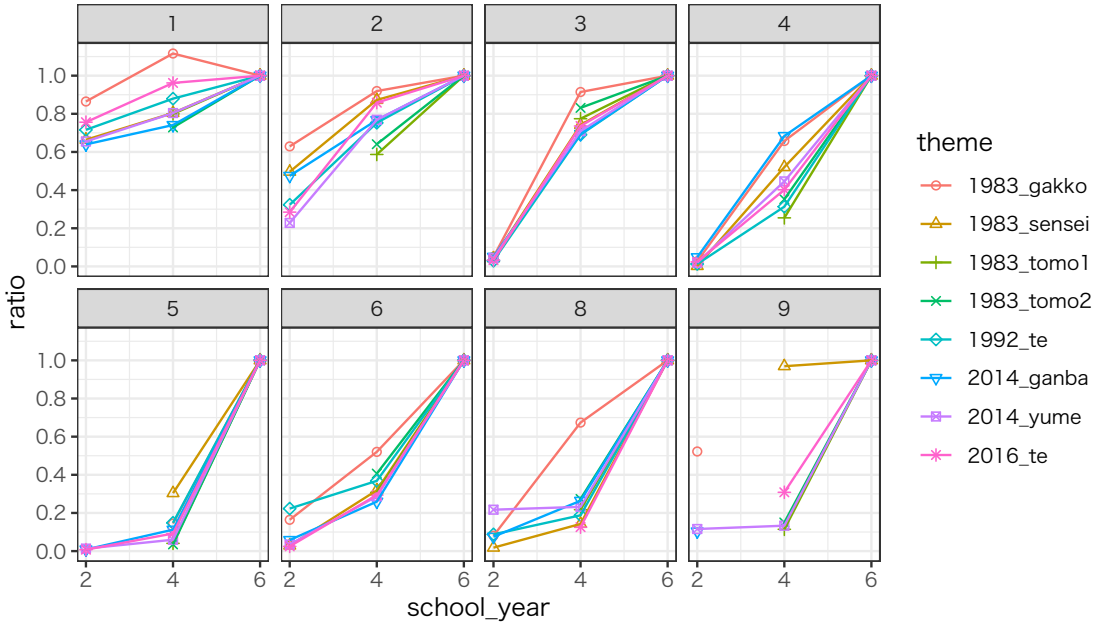

図 7　6 年次を 1 としたときの漢字使用率 (漢字配当学年別・課題別・学年別)

全体的な特徴をまとめると、2 年生や 4 年生はその学年までの配当漢字をある程度習得しているが、6 年生と同等の使用率には至っていない。例えば、2 年生は 6 年生と比べて、1 年配当漢字を 6〜8 割、2 年配当漢字を 2〜6 割の使用率で使う。同様に、4 年生は 6 年生と比べて、3 年配当漢字を 7〜9 割、4 年配当漢字を 2〜7 割の使用率で使う。しかし、その学年で配当漢字を完全に使いこなすようになるわけではなく、6 年生までの学習を通じて使用率は上昇していく。

一方、その学年以降に配当されている漢字の使用率は非常に低い。例えば、2 年生は 3 年配当以降の漢字をあまり使わず、4 年生は 5 年配当以降の漢字をあまり使わない。しかし、まったく使わないわけではない。例えば、2 年生は 3〜5 年配当漢字をほとんど使わないが、1983 年の「先生」と 1992 年の「手」では 6 年生の 2 割程度の率で 6 年配当漢字を使用している。4 年生は 5 年配当漢字を 6 年生の 0〜2 割程度使用するし、6 年配当漢字は 2〜5 割と未習としてはかなり高い率で使用する。

課題ごとの漢字使用率の違いや、未習漢字の使用は、使用される語彙の影響が大きい。6 年配当漢字の「私」はどの課題でも履修年次 (6 年) より早く使用される傾向があり、4 年生における 6 年配当漢字の使用率を押し上げる要因になっている。1983 年「学校」の 4 年生で常用漢字の使用率が高いのは、調査校所在地と関係する「松」が未習でも使用されるためのようである。1983 年「先生」の 4 年生で 5 年配当漢字の使用率がやや高いのは、授業の「授」の頻度が高いためと思われる。課題と関連する漢字は使用頻度が高くなる傾向があるが、履修年次よりも早く使われるかどうかは別の問題である。例えば 2014〜2016 年の「夢」は、6 年生では 5

年配当の「夢」という漢字がよく使われているが、4 年生ではあまり使われておらず、5 年配当漢字全体の使用率も他の課題と比べて特に高くはない。

5. おわりに

1983 年の作文コーパスを構築し、その概要を報告した。また文字種の構成比を中心として、既存コーパスとの違いを述べた。全体としてはどのコーパスも平仮名が 6 割以上を占め、ついで漢字、記号が多く、片仮名が少ないという構成比は共通しているが、「ともだち」は仮字が多く、「夢」「頑張ったこと」は外来語に由来する片仮名が多いなど、作文テーマに依拠する語彙や文字種の違いも見られた。今後、作文テーマの影響を考慮した学年別使用語彙の分析など、資料を活用した研究を進めたい。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「語彙使用・漢字使用に着目した児童の文章作成能力の経年変化の実態調査」(代表: 宮城信) の一環として実施したものである。また、科研費基盤 (B)「作文を支援する語彙・文法事項に関する研究」による成果「児童・生徒作文コーパス」(代表: 矢澤真人) および博報財団第 11 回児童教育実践についての研究助成「児童・生徒の作文能力の経年変化の解明と現場と協働した指導法の開発」(代表: 宮城信) による成果「「手」作文コーパス」を利用して行われたものである。

文　献

宮城信・今田水穂 (2018).「『児童・生徒作文コーパス』を用いた漢字使用能力の発達過程の分析」　計量国語学, 31:5, pp. 352–369.

阿部藤子・今田水穂・宗我部義則・冨士原紀絵・松崎史周・宮城信 (2017).「児童生徒の「手」作文に於ける経年変化の計量的分析: 1992 年と 2016 年の作文を比較して」　『言語資源活用ワークショップ 2016 発表論文集』　pp. 234–247.

成田信子・宗我部義則・田中美也子 (1995).「作文能力発達に関する縦断的研究 その一: 小学生から大学生に至る同題作文の分析」　国語科教育, 42, pp. 183–192.

島村直己 (1987).「児童の漢字使用: 課題作文の漢字含有率から」　『研究報告集 8 (国立国語研究所報告 90)』　pp. 77–94.