小規模コーパスの必要性と可能性

――『高校生日本語作文コーパス』の構築と利用から――

松本理美

一．はじめに

現在、日本の高校には多様な言語背景を持つ高校生が通学している。本稿では、現在日本の高校に通っている高校生をその母語や日本語習得などの言語的な背景により、日本人高校生、留学生、外国ルーツ高校生の三つのグループに分類する。外国ルーツ高校生とは、国籍、母語、家庭での使用言語は多様で、両親あるいはそのいずれかの母語が日本語ではない、日本語以外の言語が使用されるコミュニティーで育ち日本語以外の言語で教育を受けて来たなど、外国につながりを持つ生徒で、親の都合により親に伴って来日した生徒である。母語のいかんにかかわらず、生活の中で日本語以外の言語が用いられており、日本語の獲得（習得）に影響を与えているという点で日本人高校生と区別する。また、本人の意志とは関係なく来日し、段階的な日本語教育を受けることなく、日本語で教科学習をしなければならないという点で留学生とは区別する。

日本の高校における留学生と外国ルーツ高校生の中には、学習言語をはじめ日々の授業内容を理解できるだけの日本語リテラシーを持たない生徒が少なからず存在する。留学生は、留学プログラムに日本語教育が組み込まれている場合が多く、段階的、体系的な日本語教育を受けながら、他の教科の授業を受けている。しかし、外国ルーツ高校生については、日本語習得の程度は非常に多様である。「日本語指導が必要な生徒」とされていても、高校側の受入体制の問題から日本語指導が受けられず、全く日本語が理解できないまま、教室にただ座っている生徒も少なくない（二節で詳説する）。

昨今、このような外国ルーツ児童生徒が増加しているという現状が新聞等メディアに取り上げられる機会も増え、これまで国公立高校に限られていた「日本語指導が必要な児童生徒」の数的把握を主とした文部科学省調査も、外部の研究機関の協力の下で私立高校に拡大して行われるようになってきた（齋藤他、二〇二

一）。しかし、外国ルーツ高校生における日本語リテラシーの実態調査には至っていない。

外国ルーツ高校生の中には、流暢に日本語を話す生徒も多いが、日本語の読み書きに困難を抱えている生徒が少なくない。しかし、そのような生徒は「日本語指導が必要な生徒」に数えられないことが多く、高校の現場では、「書けないのは日本人高校生も同じ」と認識されることもある。阿部他（二〇二〇）が述べているように、日本人高校生についても「作文が書けない」生徒が多いことが問題になってはいるが、日本人高校生の「書けない」と外国ルーツ高校生や留学生の「書けない」は同じなのであろうか。これを調査した先行研究は極めて稀少で、分析するための公開データもないのが現状である。

そこで、本稿では、筆者が構築した『高校生日本語作文コーパス』を利用して、日本人高校生、留学生、外国ルーツ高校生の作文における書き言葉の違いを明らかにすることを目的とする。そして、小規模なコーパスではあるが、これを構築しなければ明らかにならなかった結果を基に、小規模コーパスの必要性と可能性を主張する。

以下、第二節で外国ルーツ高校生の現状を説明し、第三節でコーパスとマイノリティー言語について言及する。第四節で『高校生日本語作文コーパス』を利用した調査と分析の結果について詳細を述べ、第五節で全体をまとめる。

二．外国ルーツ高校生の現状

外国ルーツ高校生においては、日本語を含む複数の言語的な背景を持っていること、本人の意志とは関係なく親に伴って来日したことの二点のみが共通しており、それ以外は、母語、国籍だけではなく、生まれた国、親の母語、日常使用言語、初等教育での使用言語、来日時期、来日後の日本語支援の環境など、あらゆる背景が多様である。外国ルーツ高校生の親が来日した理由は、仕事によるもの、国際結婚によるものが大半を占める。

数的把握を主とした調査については、文部科学省により毎年行われている学校基本調査に加え、平成三年より二年ごとに「日本語指導が必要な児童生徒」に関する調査が、国公立高校を対象に行われている。「「日本語指導が必要な児童生徒」とは、日本語で日常会話が十分にできない児童生徒、もしくは、日常会話ができても学年相当の学習言語が不足し、学習活動への参加に支障が生じている児童生徒を指す」（文部科学省　二〇二二：一）。

文部科学省（二〇二二）による日本語指導が必要な児童生徒数の推移を外国籍の児童生徒（図1）と日本国籍の児童生徒（図2）に分けて示す[1]。

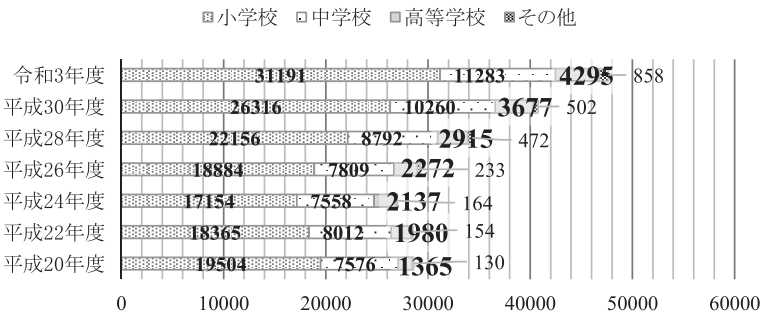

図1　日本語指導が必要な外国籍の児童生徒数

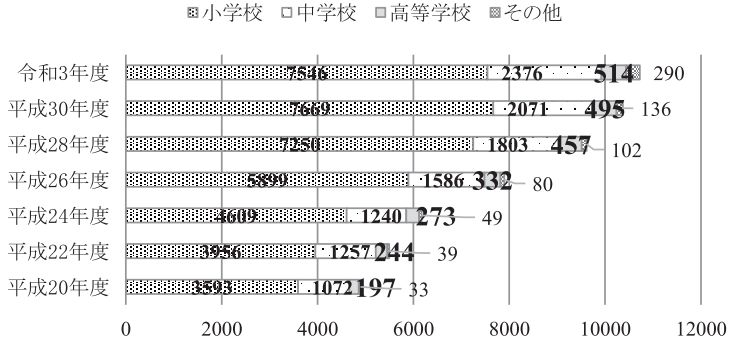

図2　日本語指導が必要な日本国籍の児童生徒数

図1、図2を見ると、日本語指導が必要な児童生徒は国籍にかかわらず年々増加しているが、文部科学省による「日本語指導が必要な児童生徒」に関する調査には問題があり、特に高校においては、「日本語指導が必要な生徒」の実態を把握しているとは言い難い。主な問題点は以下の三点である。

①国公立の学校のみを対象とした調査であり、私立学校が多い高校においては、国公立高校の調査だけでは不十分である。

②「日本語指導が必要」か否かの判断基準が設けられておらず、アンケートに応える教員の「印象」に任されている。

③留学生か否かの区別がなく、「日本語指導が必要な生徒」の言語的な背景が不明である。

二〇二一年度に文部科学省の「高等学校における日本語指導体制整備事業」により、東京学芸大学が調査委託を受け、国内の私立高校を含むすべての高校を対象にアンケート調査を行った。齋藤他（二〇二二）の報告に基づき、その一部を表1に整理した（結果はアンケート調査に応じた高校（一五九〇校）のうち、外国人生徒等[2]が在籍する学校のみ八八〇校の実数である）。なお、学校基本調査による生徒数の全数を（　）内に示した。

表1　高校における「日本語指導が必要な生徒」の調査結果

	学校数	全生徒数	外国人生徒等の数	日本語指導が必要な生徒数	日本語指導を受けている生徒数
国公立	755	353,622（1,997,541）	7,617	4,387	2,694
私立	125	91,528（1,010,631）	2,347		

齋藤他（2022）の調査結果に基づく

表1の通り、外国人生徒等の全生徒に占める割合は、国公立高校で約二%、私立高校で二・五%に過ぎず、「日本語指導が必要な生徒」は外国人生徒等の四四%であるが、全生徒数に占める割合は一%を下回る。

「日本語指導が必要な生徒」は、日本の高校に通う高校生の中では、マイノリティーの外国ルーツ高校生の中の、さらにマイノリティーな存在である。そのため、これまでは話題に上ることも少なく、実態調査も行われていなかった。しかし、昨今、「日本語指導が必要な高校生」の中途退学や進学・就職などにおける問題も顕在化しており、日本における外国人労働者や国際結婚の増加、技能実習生制度の拡充により、今後もこの人数が増えることが予想されている。ダイバーシティーやSDGsの取り組みの進展など、経済的社会的要請の変化とも併せて、「日本語指導が必要な生徒」の日本語リテラシーについては、看過できない問題になっている。

三、コーパスとマイノリティー言語

外国ルーツ高校生を巡る問題については、社会学、教育学、応用言語学など非常に学際的な研究の対象であるが、本稿では、その中の日本語リテラシー、特に作文に見られる書き言葉に着目し、収集した高校生作文をデータとしたコーパスを利用し、日本語学の視点により分析を行う。

本来コーパスとは、「言語に対する統計的なアプローチのために構築された研究装置」であり、「ある言語の研究のために、その言語で実際に用いられた用例を大量に偏りなく収集して電子化し、検索用情報を付加したもの」（前川、二〇一三）であると定義される。コーパスに求められる要件として、前川（二〇一三）は、代表性、均衡性、規模、真正性、電子化、公開、アノテーションなどを挙げているが、これらの要件を満たすコーパスを構築するには、莫大な時間と費用がかかることは想像に難くない。

日本においては、国立国語研究所が構築した『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）[3]、山崎・前川　二〇一四）に代表されるように、テキストデータを中心としたコーパスが主流であった。昨今はICT技術の急進により、テキストデータに加え、音声、画像、動画なども対象とするコーパス構築が可能とな

り、データの収集やアノテーションにおいても進歩し続けている技術を駆使した自動化が進んでいる。今では多種多様なコーパスの開発が進展し、コーパスを利用した様々な分野の研究や教育への応用が進められており、その可能性は広がり続けている。

一方で、研究対象の絶対数が少ない、データの収集に制約があり大量収集が困難である、アノテーションが自動化できないというように、ICT 技術進歩の恩恵を受けられない分野も存在し、そのような分野においては、前述した要件を満たす大規模なコーパスの構築は容易ではない。しかし、消滅の危機に瀕する少数言語や、特殊な背景を持つマイノリティーの言語が存在することも事実であり、大量データの収集が困難であってもコーパス構築の必要性が否定されるものではない。

以下、日本人高校生、留学生、外国ルーツ高校生の作文の複文に着目し、それぞれの日本語における文法的特徴を捉えるために、『高校生日本語作文コーパス』を利用した調査と分析結果について述べる。

なお、『高校生日本語作文コーパス』を構成する外国ルーツ高校生作文データは、主に「日本語指導が必要な生徒」の日本語作文を収集したものであり、四節・五節で言うところの外国ルーツ高校生とは「日本語指導が必要な」外国ルーツ高校生である。

四.『高校生日本語作文コーパス』を利用した調査と分析結果

(1)『高校生日本語作文コーパス』のデータについて

日本人高校生、留学生、外国ルーツ高校生ともに、表2の通りにコントロールされた日本語作文を収集し、コーパスデータとした。

表 2　高校生日本語作文データの内容

ジャンル	出来事についての自由作文（出来事作文）
テーマ	学校行事（体育祭、文化祭）について
作成様式	400 字詰め原稿用紙に自書、作成時間 40 分
作成環境	国語の授業中に作成（辞書使用不可、教員による添削・指導はなし）
使用言語	日本語
データ様式	テキスト（原文、修正版）、Excel（修正版の形態素解析データ）

松本（2021）に基づく

表3　高校生日本語作文のデータサイズ（グループ別）

	ファイル数	文数	1文あたりの文字数	語数		R
日本人高校生	167	1537	36.7	異なり 延べ	1754 32438	9.7
留学生	67	975	22.6	異なり 延べ	1338 12403	12.0
外国ルーツ高校生	48	669	24.3	異なり 延べ	1024 9267	10.6

収集したデータについて、グループ別データサイズを示したものが表3である。以下、本稿に用いた計量データはすべて松本（二〇二一）に基づくものである。

R値[4]とは、語彙の豊富さを表す値で、異なり語数を延べ語数の平方根で除して求めたものである。この数値が大きいほど、作文中の語彙が豊富であると言える。最も長い文を産出している、日本語を母語とする日本人高校生のR値が最低であるのは、学校行事について同様の語を用いて同様の内容の作文を書いたことによるものであると考察される。

表4　高校生作文における単文数・複文数・複文率

	単文	複文	複文率
作文（日）	272	1265	0.82
作文（留）	550	425	0.44
作文（外）	304	362	0.54

(2) 複文の計量分析の結果

次に、作文中の文に着目して計量し、グループごとの複文率を求め（表4）、複文については、その複文を構成する接続節の数を求めた（図3）。以下、図表において、日本人高校生作文、留学生作文、外国ルーツ高校生作文を、それぞれ、作文（日）、作文（留）、作文（外）など、適宜略して示すことがある。

表4をみると、日本人高校生作文では、外国ルーツ高校生作文、留学生作文に比べ、複文率が非常に高いことがわかる。複文をさらに詳細に分析し、それぞれの複文にいくつ接続節が含まれているかを計量し、グラフに表したものが図3である（グラフ内の数字は%を表す）。

図3から読み取れる特徴としては、複文を構成する接続節の数は日本人高校生作文が最も多く、日本語を母語とする日本人高校生は、単文よりも接続節を含んだ複文により文章を構成しているといえる。外国ルーツ高校生の作文では割合は高くないものの、接続節を六節以上含む文もみられた。これは外国ルーツ高校生が話し言葉のような文体で作文を書いていることを示唆するものである。なお、複文率は文章レベルに必ずしも比例するものではなく、接続節の数や文の長さのみで、作文力が測れないことは言う

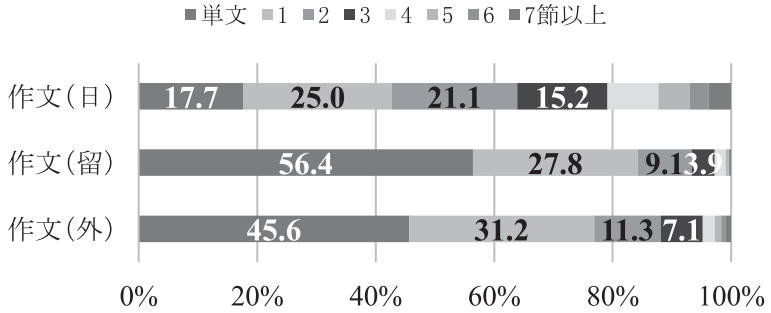

図3　高校生作文における複文を構成する接続節数の割合

表5　高校生作文における接続節の機能別出現頻度

	補足節	連体節	連用節
作文（日）	803（23.1）	695（20.0）	1985（57.0）
作文（留）	142（20.9）	140（20.6）	398（58.5）
作文（外）	127（19.2）	90（13.6）	444（67.2）

までもない。

(3) 接続節の機能別出現頻度

次に、接続節を文法的な働きにより、補足節（述語の補足語としての働きをもつ接続節）、連体節（名詞を修飾する働きをもつ接続節）、連用節（述語、または主節全体の修飾語としての働きをもつ副詞節および主節と対等の関係をもつ並列節）に分類して出現頻度を計量した。結果は表5の通りである。（　）内は出現割合（%）である。

表5からは、いずれのグループの作文においても、連体節の出現割合が低く、連用節の出現割合が高いことがわかるが、中でもその傾向は外国ルーツ高校生作文に顕著である。国語教科書に出現する接続節を調査した松本（二〇二一）では、いずれの学校種の教科書においても連体節の割合が高く、学校種が上がるほど、顕著に連体節の割合が増加し、連用節の割合が低下するという結果を得た。つまり、これは文章レベルが上がるほど、連体節の割合が増加し、連用節の割合が減少することを示唆していると考えると、特に連体節の割合が非常に低い外国ルーツ高校生作文は、文法的には稚拙である可能性が高い。

そこで連体節についてさらに詳細な分析を行う。連体節については、連体節と、連体節が修飾する主名詞（被修飾名詞）の関係により、補足語修飾節（主名詞が連体節の述語の補足語となっているもの）、命題補充節（連体節が主名詞の説明となっているも

の）、相対名詞節（補足語修飾節、命題補充節で定義できないもの）の三つのタイプに分けて計量分析を行った。日本人高校生作文、留学生作文、外国ルーツ高校生作文におけるタイプ別出現割合を図4に示した。

いずれのグループの高校生作文においても補足語修飾節の出現割合が最も高いが、外国ルーツ高校生作文と留学生作文にはこれが特に高い割合で出現することが明らかである。命題補充節については、一〇〇文あたりの調整頻度を図5に示した。

図5からは、外国ルーツ高校生作文、留学生作文における命題補充節の出現頻度は、日本人高校生作文の三割にも満たないことがわかる。松本（二〇二一）によると、国語教科書においても、すべての学校種で補足語修飾節が過半数を占めており、割合だけで言うと、中学校教科書が日本人高校生作文と同程度、小学校教科書が留学生作文と同程度となっている。しかしながら、命題補充節の一〇〇文あたりの調整頻度を見ると、いずれのグループの高校生作文も小学校教科書より少なく、極めて低頻度であることがわかった。

さらに、外国ルーツ高校生作文には、連体節で表現できるものでも、難易度の低い連用節で表現する傾向が見られ、これも連体節の割合が低く、連用節の割合が高くなる原因の一つであると考察できる。

図6は、命題補充節が修飾する主名詞を、日本語能力試験の語彙レベルを援用して分類したものである。一級が最も高いレベル

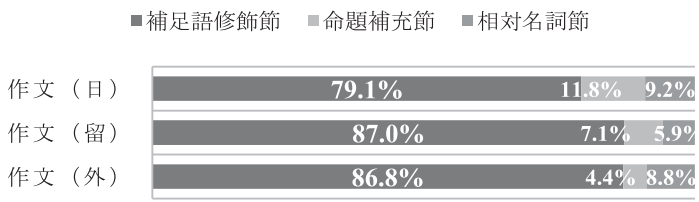

図4　高校生日本語作文における連体節のタイプ別出現割合

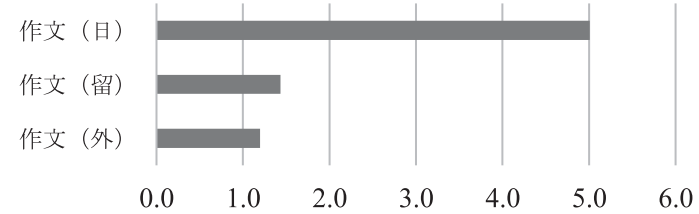

図5　高校生日本語作文における100文あたりの命題補充節の頻度

で、数字が大きくなるほど語の難易度が下がる。

図6からは、留学生が最も難易度の高い語を主名詞に用いており、前述した語彙の豊富さを示すR値の高さと併せて考えると、留学生の語彙力が最も高いことが示唆される。全体として命題補充節の主名詞となった語は、「感じ」「姿」や、「競技」「種目」「練習」など、作文のテーマに関連すると思われる語が多く用いられていることがわかった。

最後に、連用節を用法ごとに分類し、出現頻度について計量分析を行う。用法ごとの出現割合をグループ別にグラフ化したものが図7である。

図7の通り、日本人高校生作文には時間節の割合が低く、原因理由節の割合が高い。留学生作文には時間節と逆接節の割合が高く、外国ルーツ高校生作文には時間節の割合が特に高いことが確認された。

節末形式のバリエーションについては、紙幅の関係で、出現割合について用法ごとに特徴のみを述べる。《時間節》では、出現割合が最も低い日本人高校生作文においてバリエーションが非常に豊富であった。外国ルーツ高校生作文にテ形が六割近い割合で出現する一方で、留学生では二割に満たないという特徴がみられた。《原因理由節》においては、外国ルーツ高校生作文にはテ形が五割弱と高い割合で出現しているが、日本人高校生作文、留学生作文でも四割弱と出現割合は高い傾向にあった。《条件譲歩節》については、外国ルーツ高校生作文において、話し言葉の節末形

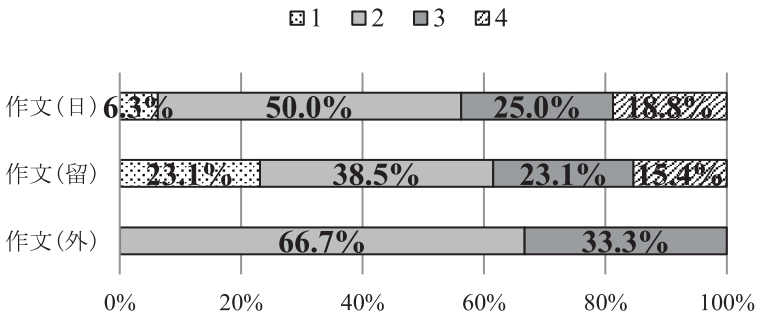

図6　高校生日本語作文における命題補充節主名詞の難易度

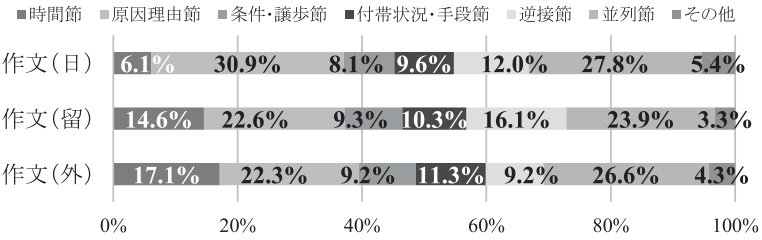

図7　高校生作文における用法別連用節の出現割合

式に多いとされる「たら」が最も高い割合で出現したのに対し、留学生作文では「ば」の出現割合が「と」に続いて高かった。《付帯状況節》ではいずれのグループの作文においても、テ形の出現割合が七割以上を占めていた。《逆接節》の節末形式については、「けど」の出現割合がいずれのグループにも高い割合で出現しているが、外国ルーツ高校生作文には六三％と顕著に高い割合で出現していた。一方、「が」「けれど」の合計は、留学生作文での出現割合が最高であった。

第四節で述べた分析結果について、語彙、文法、産出量の三つの観点で考察する。《語彙》については、語彙の豊富さを表すR値と語の難易度から、留学生の作文における語彙が最も豊富で、難易度が高いことが明らかとなった。これは、留学生はすべて漢字圏（中国出身）の生徒であり、母語による語彙力が転移しやすいことによると考えられる。《文法》については、日本人高校生作文において、連体節、中でも命題補充節の割合が他の二つのグループに比べ高かったことは、文法的に複雑な文の産出においては、日本語を母語とする日本人高校生が有利であること示している。留学生作文においては、連体節が日本人高校生作文と同程度の割合で出現しており、語彙力とともに母語による作文力が転移していることが推測される。《産出量》については、文の長さ、複文率、一文に含まれる接続節数のいずれにおいても、日本語を母語とする日本人高校生作文で突出していることが確認できた。

5. まとめ

本稿では、日本の高校に通う高校生を言語的な背景により、日本人高校生、留学生、外国ルーツ高校生の三つのグループに分け、その作文について『高校生日本語作文コーパス』を用いて調査、分析を行った結果について述べた。

高校生日本語作文の調査、分析の結果、それぞれのグループの作文の特徴は以下のようにまとめられる。

【日本人高校生作文】

- 作文中の一文の長さ、一文に含まれる接続節数は、三つのグループの中で群を抜いて多く、複文率は高い。
- 接続節の中では、連体節の出現割合と、連体節の中では複雑であると考えられる命題補充節の出現割合が三つのグループの中で最も高い。
- 連用節の節末形式でいうと、時間節のバリエーションが群を抜いて多く、また作文のテーマから出現率が高いことが予想される原因理由節が、高い割合で出現している。

【留学生作文】

- 語彙の豊富さを表すR値、語彙の難易度については、いずれも最も高い値であり、母語の語彙力の高さが示唆される結果となった。
- 連体節については、命題補充節より難易度の低い補足語修飾節

の割合が高いものの、連体節の割合が日本人高校生作文と同程度の割合で出現する。

- 連用節についても、話し言葉に多いとされる「たら」より「ば」が高い割合で出現している、話し言葉で用いられる「けど」の出現割合が三つのグループの中で最も低いことなど、体系的な日本語教育により書き言葉を習得していることが示唆される。

【外国ルーツ高校生作文】

- 産出量については、日本人高校生作文に次いで二番目であり、日本人高校生作文同様に、長い文（接続節を多く含む文）を産出できることがわかる。
- 一方で、連体節の出現割合は最も低く、連用節の出現割合は最も高い。
- 連用節の節末形式では、時間節、原因理由節、付帯状況節のいずれにおいてもテ形節を多く用いており、逆接節では「けど」を多用していることなどから、丸山（二〇一四）が指摘するとおり、文体に話し言葉的特徴があることが確認できた。

以上より、留学生作文には、日本語書き言葉を体系的、段階的に学習していることを裏付ける特徴が見られ、母語の影響もあり、高い語彙力を示唆する結果が得られた。一方で、外国ルーツ高校生には、産出量、文法（連体節、連用節）のすべてにおいて、話し言葉から日本語を習得しており、書き言葉については段階的、体系的な日本語教育が行われていないことを裏付ける結果が得られた。外国ルーツ高校生の八割以上が留学生同様に中国語母語話者であることを考えると、言語習得・言語教育の視点において非常に興味深い結果であると言える。

六万語足らずの非常に規模の小さなコーパスであっても、このような言語的特徴を捉えることができ、今後の日本語教育や国語教育の進展に資する可能性を示した。それぞれの研究目的に合ったデータ収集とアノテーションにより構築された小規模コーパスがマイノリティー言語の研究には必要であり、小規模コーパスの構築とその利用によって、大規模コーパスで取りこぼした言語的特徴を発見し得る可能性は大きいと言えよう。

注

（1）図1・図2は、文部科学省（二〇二二）に基づく。図中の「その他」は、義務教育学校・中等教育学校・特別支援学校の生徒数を合計したものである。

（2）外国籍の生徒に加え、外国にルーツがある日本国籍の生徒を含んでいる。

（3）Balanced Corpus of Contemporary Written Japanese（Maekawa et al., 2014）の略称。

（4）延べ語数の影響を緩和するため、R（guiraud 指数）を用いる。

付記

本稿は、第六六回立命館大学日本文学会大会（二〇二二年六月一二日　オンライン開催）での口頭発表「小規模コーパスの必要性と可能性―高校生日本語作文コーパスの構築を通じて―」をもとに、加筆修正を行なったものです。発表に際し、貴重なご意見・ご助言を賜りました。ここに記してお礼申し上げます。

参考文献

安部朋世（二〇二〇）「国語教育・初年次教育等に関する近年の状況」森山卓郎・矢澤真人・安部朋世「言語習熟論へ向けて―日本語研究と国語教育・初年次教育など―」日本語学会二〇二〇年度秋季大会予稿集　二二七―二三六頁　日本語学会.

齋藤ひろみ・武内博子・南浦涼介（二〇一二）「高等学校における外国人生徒等への日本語指導の現状と課題―質問紙調査から―」二〇一二年度日本語教育学会春季大会予稿集　二四三―二四八頁　日本語教育学会.

前川喜久雄（二〇一三）「第一章　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』一―三一頁、朝倉書店.

松本理美（二〇二一）「外国ルーツ高校生を含む『高校生日本語作文コーパス』の構築と計量的研究」未刊行博士論文　立命館大学.

丸山岳彦（二〇一四）「現代日本語の多重的な節連鎖構造について―CSJ と BCCWJ を用いた分析―」石黒圭・橋本行洋編『話し言葉と書き言葉の接点』九三―一一四頁、ひつじ書房.

文部科学省（二〇二二）「日本語指導が必要な児童生徒の受入状況等に関する調査（令和三年度）」の結果について、http//www.mext.go.jp/b_menu/houdou/31/09/1421569_00003.htm（二〇二三年一月三一日参照）.

山崎誠・前川喜久雄（二〇一四）「第1章　コーパスの設計」前川喜久雄監修　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』一―二一頁、朝倉書店.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka and Yasuharu Den（二〇一四）*Balanced Corpus of Contemporary Written Japanese*. Language Resources and Evaluation（LRE 48）. 三四五―三七一頁.

（まつもと・さとみ　大阪樟蔭女子大学特任准教授）