

(特別寄稿)

「わたしはマイク・ミラーです」を再考する

— 日本語コーパスの教育応用をめぐる —

石川 慎一郎

1. はじめに

コーパスが外国語教育を変革する、という可能性が初めて具体的に示されたのは1987年のことであった。この年、英国バーミンガム大学の John Sinclair 氏が、英国の辞書編纂の歴史を変える、全面的なコーパス調査に基づく革新的な辞書、*Collins COBUILD English Language Dictionary* を出版したのである。それまでの辞書の記述内容を大きく踏み出し、母語話者ですら気づいていなかった語の意味の隠れたニュアンスや、語と語の典型的な結びつきを明らかにし、それらをコーパスから得られた生きた用例とともに学習者向けにわかりやすく提示したコーパス準拠辞書の出版は、経験を積んだ母語話者教師（専門家）の直観と経験だけに依拠する言語教育の限界と、その限界を補う有効な補助具としてのコーパスの可能性を世に示したのであった（石川、2012、p.166）。

それから35年が経過し、多くの英語コーパスが開発・公開され、いまや、英国だけでなく日本においても、コーパスを利用しない英語辞書を探すほうが難しい。また、英語母語話者の L1 産出を集めたコーパス（母語話者コーパス）に加え、近年では、学習者の L2 産出、つまりは、「中間言語」を集めたコーパス（学習者コーパス）の構築も進み（石川、2017、pp.64-69）、学習者が典型的に犯しがちな誤りを特定し、それが自然に克服できるよう工夫された教材（たとえば、ケンブリッジ大学出版局の *Touchstone* シリーズ）も出版されている。また、近刊の和英辞典（岸野（編）、2019）においては、学習者コーパスと母語話者コーパスを比較して、日本人英語学習者が過剰・過少使用しがちな語を具体的に特定し、ユーザーに注意喚起するユニークなコラムが設けられている。いまや、コーパスは、英語の研究はもちろん、英語教育のあらゆる分野に浸透したと言える。

翻って日本語のことを考えると、コーパス活用はまだ緒に就いたばかりである。もっとも、研究が普及する前提となる本格的な日本語コーパスが公開されたのは2011年のことであった。すでに1960年代に本格的なコーパスがリリースされていた英語と比べると、日本語のコーパス研究が約50年遅れのス

タートであったことは考慮されなければならない。現在、日本語の研究者や日本語教育の関係者は精力的な活動によって、このずれを一気に縮めようとしている。

では、コーパスとはどのようなもので、それを使うことで日本語教育の何がどのように変わるのであろうか。以下、2章では、日本語コーパスの代表格である「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese: 以後 BCCWJ) について概観する。その後、3章では、身近な実例として、代表的な日本語教科書である『みんなの日本語(初級・第2版)』(スリーエーネットワーク、2012) の第1課に登場する「わたしはマイク・ミラーです」という文を取り上げ、BCCWJを調査することによって、この1文に関して、どのような言語事実がコーパスから引き出せるかを例証する。

2. 「現代日本語書き言葉均衡コーパス」の概要

日本語の母語話者コーパスの代表格である BCCWJ は国立国語研究所によって開発された。英語コーパスの代表格である British National Corpus (BNC) が英国政府の資金援助で作られたという背景を持つように、コーパスを構築し、自言語を1つの形として永遠にとどめようとする仕事は時として国家的な公共性を帯びる。

コーパスでは、母集団の多様性を反映する形で均衡的にデータ収集を行うことが理想とされる。たとえば、書き言葉だけでも、新聞・雑誌・書籍・各種パンフレット等、様々なものが存在し、それぞれ、使用される語彙や構文が異なっている。ゆえに、仮に(比較的収集が容易な)新聞データだけでコーパスを作ったとしても、それは「新聞の日本語」を代表するものではあっても、「現代日本語の書き言葉全体」を代表するものにはなりえない。

そこで、BCCWJの開発チームは、以下のように、多様なソースからデータを収集して1億語の大型コーパスを構築した(表1)。表1に記載のデータは BCCWJ の検索サイトである「少納言」上の解説による。なお、BCCWJ の開発チームは、データの種類について、「少納言」では「メディア/ジャンル」という語を、中上級者用の検索サイトである「中納言」では「レジスター」と「ジャンル」という語を用いているが、本稿では、より一般的な「テキストタイプ」という用語を用いる。

表1 BCCWJの構成

テキストタイプ	年代	サンプル件数	語数(概算)
書籍	1971-2005	22,058	6,270 万
雑誌	2001-2005	1,996	440 万
新聞	2001-2005	1,473	140 万
白書	1976-2005	1,500	490 万
教科書	2005-2007	412	90 万
広報紙	2008	354	380 万
Yahoo! 知恵袋	2005	91,445	1,030 万
Yahoo! ブログ	2008	52,680	1,020 万
韻文	1980-2005	252	20 万
法律	1976-2005	346	110 万
国会会議録	1976-2005	159	510 万

この中には、校閲を経たもの(書籍・雑誌・新聞・白書・教科書・広報紙・法律等)もあれば無校閲のものもあり、創作(書籍の中の「小説」・韻文)もあれば事実の紹介を行うものもある。また、限定的な読者を対象としたもの(教科書・法律)もあれば幅広い読者に向けたものもあり、相手との双方向のやりとりを念頭に置いたもの(Yahoo! 知恵袋の質問と回答、国会会議録の質疑)もあればそうでないものもある。さらに、高度に専門的な内容を含みうるもの(白書・法律・国会会議録・Yahoo! 知恵袋・Yahoo! ブログ)もあれば一般的な内容のものもある。このように、1種や2種のテキストタイプだけでコーパスを構築するのではなく、言語としての性質が異なる11種類のテキストタイプを包含していることが、言語研究資料としてのBCCWJの価値の源泉となっている。

もっとも、BCCWJの使用にあたっては注意すべき点もある(李・石川・砂川、2018、pp. 31-32)。というのも、BCCWJの正式名称は「現代日本語書き言葉均衡コーパス」であるが、「現代日本語」「書き言葉」「均衡」という3つの基本概念の各々について、一定の前提が存在するからである。

まず、「現代日本語」という点についてであるが、BCCWJのデータは1971年から2008年までの範囲で集められており、そこには、約40年間の幅が存在する。たとえば1970年代や1980年代によく使われていた語や表現が、その後、ほとんど使われなくなっているという可能性もあり、「現代日本語」のコーパ

スに出ているからと言って、そのすべてが現在の標準的な日本語の使い方であるとは限らない。また、テキストタイプによって、収録年数のずれも存在する。さらに、現代の出版物であっても、歴史書や歴史小説では、古語や古文が含まれることもある。

次に、「書き言葉」という点についてであるが、BCCWJには、「書き言葉」という範疇の周縁に存在するテキストも含まれている。たとえば、(A)話し言葉をイメージして書かれた書き言葉（書籍中の小説に含まれる登場人物のセリフや、新聞や雑誌における他者の発言の要約的な引用部分）、(B)（一部は原稿に基づいてなされた）実際の話し言葉をさらに書き起こして加工を加えた書き言葉（国会会議録）、(C)高度に対話的で話し言葉性を有した書き言葉（読者に話しかけるようなくだけた文体で書かれたYahoo! ブログの一部等）等である。本来、書き言葉と話し言葉の境界は必ずしも明確なものではなく、それはBCCWJにもあてはまる。

最後に、「均衡」という点についてであるが、BCCWJには、大部分の書籍のように、母集団からの統計的サンプリングで集められたデータと、Yahoo!の知恵袋やブログ、また、法律や韻文のように、一定の範囲の中で入手できたデータを全量集めているものが混在している。つまり、統計的な均衡性が確保されているテキストタイプも存在するが（ただし、1サンプルの長さはそろえられていないものもあるので注意が必要）、コーパス全体をひと固まりとして見た場合に母集団としての日本語の多様性が均衡的に反映されているというわけではない（たとえば、BCCWJに含まれる新聞データは140万語、広報紙データは380万語だが、現実の日本語において新聞のテキスト量と広報紙のテキスト量が厳密にこの比率になっているわけではない）。均衡性というのはコーパス言語学のキーワードの1つであるが（石川、2012、pp. 22-25）、その意味にはもともと一定の幅がある。つまり、1種またはごく限定的な数のテキストタイプだけに偏っていないという意味での広義の「均衡性」もあれば、母集団に存在する様々な下位要素の出現比率を調査し、その比率を正確に再現するという意味での狭義の「均衡性」もある。BCCWJの均衡性についてもそうした幅をもって理解することが肝要であろう。なお、BCCWJの1億語をひと固まりとして検索する際には、書籍（6,270万語）と2種のウェブテキスト（合計2,050万語）だけで全体の8割以上を占めていることから、サイズの小さいほかのテキストタイプの特性はほとんど消えていることに留意したい。

BCCWJは、DVDを入手し、各自がコーパス分析用ソフトウェアを用意して分析する方法のほか、初中級者用のオンライン検索インタフェース「少納言」を介して分析する方法、中上級者用のオンライン検索インタフェース「中納言」を介して分析する方法、BCCWJのデータの一部を使い、様々な言語分析が平易に行えるよう開発されたオンライン検索インタフェース「NINJAL-Lago Word Profiler for BCCWJ (NLB)」を介して分析する方法などがあり、ユーザーのニーズに応じて使い分けることができるようになっている。

3. 「わたしはマイク・ミラーです」のコーパス分析

3.1 ねらいとRQ

「わたしはマイク・ミラーです」という文はもちろん文法的には問題がない。しかし、コーパス言語学では、文が文法的に正しいか否かということだけでなく、その文がどういう環境で出現するか、他にどのような表現上の変種が存在するか、それらと比べて対象文が実際の言語運用の中でどの程度自然で典型的であるか、といった点を重視する。

その際、判断の根拠となるのが頻度である。たとえば類義のAとBという表現があったとして、仮にAの頻度がBの頻度より十分に高ければ、Aのほうがより典型的で自然な表現であると考えるのである。

こうした観点からこの文を改めて見た場合、いくつか確認したい点が存在する。ここでは5点を取り上げてリサーチクエスションとする。

RQ1 1人称単数代名詞として「ワタシ」はどの程度典型的か？

RQ2 「ワタシ」の表記法として「わたし」はどの程度典型的か？

RQ3 外国人の姓名の表記法として「X・Y」はどの程度典型的か？

RQ4 自分の名を名乗る自己紹介文は全てのテキストタイプに満遍なく出現するか？

RQ5 自分の名を名乗る自己紹介文の中で「私は<姓名>です」という文型はどの程度典型的か？

これらは、いずれも、一見自明なようで、実際には、母語話者であっても即答しにくいポイントである。たとえば、RQ1について言うと、1人称代名詞である「ボク」や「オレ」に対して「ワタシ」がどの程度の頻度上の優先性を持つかはわからない（ここでは表記形を統合した語彙素形をカタカナ

で示す)。また、RQ2については、教科書に出てくる「わたし」という表記が、「私」「わたくし」「あたし」「あたくし」といった他の表記に比べての程度典型的であるのかもはっきりしない。

RQ3について言うと、日本語にはカタカナ人名の表記法として「マイク・ミラー」のほかに「マイク＝ミラー」も存在する。姓名の区切りにはナカグロ(中点)、二重姓ないし二重名にはダブルハイフンを使用するという立場もあるが(小林、2015)、知る限り、この点についての公式なガイドラインは存在せず、実際、校閲されたテキストにおいて、姓名をダブルハイフンで区切る例も存在する。以下はともにBCCWJ用例であるが、イギリスの詩人 William Wordsworth の名前が2つの記号で表記されている。

- (1) ウィリアム・ワーズワースの詩の一編を口ずさみながらあの日々を偲びました。(Yahoo! ブログ)
- (2) …フランスのヴィクトル＝ユーゴー (Victor Hugo)、ウィリアム＝ワーズワース (William Wordsworth)、ウォルター＝スコット…(書籍・社会)

この事例にも示されるように、ナカグロ記号を用いた表記がダブルハイフンに比べてどの程度典型的であるのかも自明ではない。

RQ4については、「私は<姓名>です」という文型が、様々な日本語のテキストタイプにおいて均衡的に出現するの否か、といった点に疑問が残る。

最後に、RQ5については、たとえば、「私は<姓名>です」という言い方が、「私は<姓>です」や「<姓>です」といった他の言い方に対して典型性を持つのかどうか、また、この文型にほかの挿入要素がありうるのかどうかという点もはっきりしない。

以下、BCCWJの検索を通して、これらの点に一定の回答を得ることを目指す。こうした言語事実を整理し、蓄積していくことは、日本語教育で用いる文型や例文の検証にも有益であろう。

3.2 データと手法

本研究ではBCCWJを使用し、オンライン検索インタフェース「中納言」で検索を行う。なお、検索日は2018年12月5-6日である。

まず、RQ1については、中納言の品詞検索では「人称代名詞」という区切りがないことから、「代名詞」全体に対して検索を行い、150万例を超えるヒッ

ト例の中からダウンロード上限の10万例のデータを入手した。その後、手作業で1人称を指している22,775例を抽出した。

次に、RQ2については、上記で得られたもののうち、語彙素が「私」となっている15,120例を抽出し、非漢字の出現形の頻度調査を行った。なお、助詞が連結した特殊形である「わたしゃ」と「あたしゃ」はそれぞれ「わたし」と「あたし」の頻度に繰り込んだ。

RQ3では、まず、＜固有名詞＋補助記号（一般）＋固有名詞＞の連鎖で検索を行い、67,843例を得た。ここから、カタカナの「ア」で始まる4,185例を分析サンプルとして、手作業で、外国人の人名連結の例になっていると思われるものを抽出し、補助記号別の頻度を比較した。

RQ4については、＜固有名詞（姓）＋です＋句点＞および＜固有名詞（名）＋です＋句点＞の2つの形で検索を行い、前者で217例、後者で439例、延べで656例を得た。その後、人名以外の混入例（例：～もかのうです／ニューオータニです／カワイ（注：ピアノ）です）や、明らかに名乗りや自己紹介でないもの（例：～は誰ですか？－山田です／～で有名なのが岡本です）を手作業で除き、最終的に340例を得た。これらをテキストタイプ別に整理し、頻度を比較した。

RQ5については、以上の340例のうち、自己紹介や名乗りに関して特殊な形式が出現しやすいと思われる4種のテキストタイプ（連載形式で同じ自己紹介が反復されやすいブログ、質問者と回答者のやり取りの過程で同じ自己紹介が反復されやすい知恵袋、定型的な自己紹介を含む首長挨拶が掲載されやすい都道府縣市町等の広報誌、＜指名→名乗り→質問＞という定型的やりとりがなされる国会会議録）に出ているものを除き、より一般的な自己紹介が見られる書籍（図書館母集団からの抽出データ、出版母集団からの抽出データ、ベストセラーの3種を含む）に出現した144例だけを取り出し、質的に検証した。

まず、これらの144例を4種の機能タイプに区分し（一般・電話応答・読者向けメッセージ等・国会質疑）、当該文型の出現に関して特定の制約的条件がないと思われる一般型の94例の構造を詳細に調査し、文型内要素と対象文型の典型性を検証した。

3.3 結果と考察

3.3.1 RQ1 1人称代名詞としての「ワタシ」

一般に、1人称単数代名詞としては「ワタシ」や「ボク」や「オレ」が想定され、その中では「ワタシ」が一般的であるように思われるが、実際の書き言葉ではどのような1人称代名詞が使用されており、かつ、その中で「ワタシ」の典型性がどの程度であるかは不明である。

コーパス検証の結果、10万例の代名詞のうち、1人称代名詞と思われる語彙素（異表記は区別せず）は20種、合計、22,775例で、内訳は以下の通りであった。

表2 1人称単数代名詞のタイプ（語彙素）

語	頻度	構成比 (%)	語	頻度	構成比 (%)
私	15120	66.39	朕	22	0.10
僕	3815	16.75	拙者	19	0.08
俺	2546	11.18	我が輩	17	0.07
我	466	2.05	わっち	9	0.04
わし	400	1.76	こちとら	5	0.02
己	170	0.75	わて	2	0.01
余	95	0.42	余輩	2	0.01
てまえ	35	0.15	拙僧	2	0.01
わっし	24	0.11	吾	1	0.001
麻呂	24	0.11	身共	1	0.001

表2に基づく、語彙素としての「ワタシ」は予想通り1位になっているが、実際には、それ以外の1人称代名詞も累計で4割近く使われていることになる。そのうち、10%を超えるのは、「ボク」(16.75%)と「オレ」(11.18%)のみである。現代日本語の1人称代名詞として最も典型的なのは「ワタシ」であるが、「ボク」と「オレ」の2種も一定の典型性を持つと言えるだろう。荻野(2017)は、東京の大学生対象の調査結果をふまえ、最近、男性が「自分」を、女性が「うち」を1人称代名詞的に使用するようになってきていることを報告している。品詞タグに基づく今回の分析では含まれていないが、BCCWJにもこうした用例が実際に出現している。

なお、上記の中には、「麻呂」や「拙者」など、一般の現代日本語では使用

されないものが混在しているが、これは、すでに述べたように BCCWJ に時代小説などが含まれているためである。

3.3.2 RQ2 「ワタシ」の表記法としての「わたし」

「ワタシ」は一般的には「私」や「わたし」と表記されるが、実際には、「わたくし」「あたし」など、いくつかの変種も存在する。その中で、教科書に使われている「わたし」はどの程度の典型性を持つのであろうか。

語彙素が「私」となっている15,120例を検証したところ、各表記形の内訳は以下の通りであった。

表3 「ワタシ」の表記タイプ

表記形	頻度	構成比 (%)	表記形	頻度	構成比 (%)
私	11093	73.37	あたし	11	0.07
わたし	3085	20.40	ワタシ	5	0.03
あたし	718	4.75	アタシ	2	0.01
わたくし	170	1.12	アチシ	2	0.01
わたい	17	0.11	妾 (わらわ)	1	0.01
あたくし	16	0.11	—	—	—

上記より、「ワタシ」の表記については、ひらがな表記の中では「わたし」が最も一般的であるものの、実際には、漢字表記がさらにその3倍以上を占めることが示された。また、「わたし」の公式表記であると考えられる「わたくし」の構成は1%程度で、くだけた女性1人称表記である「あたし」よりもさらに頻度が低いことがあわせて示された。

3.3.3 RQ3 外国人姓名の表記法

外国人の姓名連結であると判断されるサンプルを分析した結果、連結記号はナカグロとダブルハイフンだけでなく、シングルハイフンも存在することがわかった。

- (3) 昨年の覇者ゴラン・イワニセビッチ (クロアチア) が欠場したため、第3シードのアンドレ・アガシ (米国) が登場する… (新聞)
- (4) 五賢帝最後のマルクス=アウレリウス=アントニヌス帝の治世末期こ

ろから… (教科書)

- (5)「北虜」とはアルタン・ハンに代表されるモンゴル勢力を、「南倭」とは… (書籍・歴史)

それぞれの出現頻度は以下の通りであった。なお、等記号はダブルハイフンとみなす。

表4 外国人名連結記号

記号	頻度	構成比 (%)
ナカグロ (A・B型)	3193	97.32
ダブルハイフン (A=B型)	73	2.22
ハイフン (A-B型)	15	0.46

上記より、外国人名の表記については、教科書にある「マイク・ミラー」型のナカグロでつなぐ形が事実上の標準表記になっていることが確認された(新聞や雑誌は例外なくすべてがナカグロであり、社内的な校閲ルールでナカグロに統一していると思われる)。

なお、ハイフンの15例のうち、12例は同一の出典(書籍)であり、人名表記としてのハイフンは実質上許容されないと考えたほうがよいであろう。

一方、ダブルハイフンは73例で、ある程度の出現事例が確認された。このうち、書籍が51例、教科書が8例を占める。出典には歴史書と美術書(中高の社会・美術の教科書も含む)が多く、ある程度専門的な文脈で、歴史上の有名な人物に言及する場合にダブルハイフンの使用が相対的に高まるようである。ただし、この点については、以下のような例も見つかった。

- (6)アンリ＝ルソー (『美術2・3下:美術の広がり』日本文教出版、2005)

- (7)アンリ・ルソー (『高校美術1』日本文教出版、2006)

上記は、同じ出版社の検定教科書が同一人物について異なる記号を使用しているもので、新聞や雑誌に比べると、検定教科書の編纂では、外国人名の表記法としてナカグロを使用するルールは徹底していないと思われる。

3.3.4 RQ4 自己紹介文が出現するテキストタイプ

自分の名に言及すると思われる「<姓/名>です」文型の出現頻度は、テ

キストタイプ別に以下のようになった。なお、テキストタイプごとにデータの総量に差があるので、それぞれサイズをふまえて（表1の概数を使用）100万語あたり調整頻度（PMW）に換算した値をあわせて示す。

表5 テキストタイプ別頻度

テキストタイプ	頻度	PMW
国会会議録	58	11.37
ブログ	100	9.80
知恵袋	24 (12)	2.33 (1.77)
書籍	144	2.30
雑誌	9	2.05
広報誌	4	1.05
教科書	1	0.0001

以上より、自己紹介文は様々な日本語のテキストタイプに満遍なく出現するのではなく、国会会議録とブログというそれぞれ特殊性の強いテキストタイプに特化して出現することがわかった。逆に言えば、自己紹介文は、言語教育できわめて重視される割には、実際の日本語においてそれほど広く使われるものではないことになる。下記はブログと国会会議録からとった用例である。

- (8) お忙しい中お越しいただいてありがとうございます。私は、このブログの当事者ひろゆきです。ブログでは、日々のこと、モーニング娘。のこと、ハロプロのこと…（ブログ）
- (9) 質疑を続行いたします。#松本大輔君。#○松本（大）委員#民主党の松本大輔です。水曜日の新聞報道だったと思いますが、大臣が答弁に立ったび…（国会会議録）

自己紹介文には、(8)のように、本人起点のものと、(9)のように、他者による紹介を受けた他者起点のものがある。国会会議録では後者が特に多い。

さて、これら2つのテキストジャンルに次いで自己紹介文が多く出現するのは知恵袋であるが、出現例の24例のうち、12例は同一表現（有名芸人の語り芸の話し出しのセリフ）であった。これを除くと12例となってPMWは1.77

に下がる。

つまり、国会会議録・ブログに次いで自己紹介が多く出現するテキストタイプは書籍ということになる。会議録・ブログに比べれば頻度ははるかに低いものの、書籍において、他のテキストより相対的に多く自己紹介文が出現するのは、小説中に登場人物の名乗りのシーンが多いことが関係している。

- (10) 「あ、すみません。あたし、七瀬…七瀬美雪です。それとこのひと、あたしの幼なじみで金田一」… (書籍・ベストセラー)
- (11) …おばあちゃんが、わたしの世話を下さることになった。「亜也です。よろしく願います」と小さい声で言う。… (書籍・ベストセラー)
- (12) 新井さん、とおっしゃるんでしょ？ #「新井和代です。あのー何か、手落ちがございました？… (書籍・ベストセラー)

なお、上記の(10)と(11)は本人起点、(12)は他者起点の自己紹介と思われる。

以上で概観したように、日本語の自己紹介文は、少なくとも書き言葉について言うと、幅広いテキストタイプで均衡的に出現するというものではなく、むしろ、国会会議録とブログという特殊な環境に限って多く出現する表現型である。

3.3.5 RQ5 「私は<姓名>です」文型の形態

次に、書籍中に出現した144件の「<姓/名>です」用例に絞って機能タイプを分類したところ、以下の結果を得た。

表6 「<姓/名>です」文の機能タイプ

状況	頻度	構成比 (%)
一般	94	65.3
電話応答	34	23.6
読者向けメッセージ等	15	10.4
国会質疑	1	0.7

上記のうち、電話での応答文、書籍のあとがき等での読者向けメッセージ等、国会質疑、の3種は同じ自己紹介であってもかなり特殊なものと言える。

そこで、これらを除いた一般的な事例94例について、実際の出現形態を調査したところ、以下のような形態上のバリエーションが認められた。なお、今回は、「〈姓／名〉です。」の後続要素は分析対象外とする。

[A：前部要素] + [B：1人称要素] + [C：先行形容要素] + [D：名前要素] + です。

このうち、Aには、A1：フィラー等（例：あのー、えーと、すみません、どうも）、A2：呼びかけ（例：～さん、～先生、みなさん）、A3：あいさつ（例：こんにちは、おはようございます）、A4：応答（例：[～さんですか？－] はい、そうです）の下位区分がある。Bには、代名詞と助詞の組み合わせにいくつかのパターンが認められる。また、CにはC1：動詞要素（例：～した、～している）とC2：名詞要素（例：〈所属先〉の、〈職種〉の）の区別がある。

以上の各要素はある種のスロットとして理解することができる。下記は、大半のスロットに要素が入った場合の一例である（作例）。

(13) [A1：どうも、] [A2：みなさん、] [A3：おはようございます。] [B：私は] [C1：本店から来た] [C2：営業三課の] [D：山田（太郎）] です。

この文型では、名前に関わるD要素は必須項となるが、そのほかのスロットは義務的なものではなく、すべてのスロットが未充足になったり、あるいは、一部のスロットのみが充足されたりする場合もある。

下記は、名前部分を除くA～Cのスロットについて、それぞれの充足度と具体的な充足要素を示したものである。

表7 スロットごとの充足度と充足要素

スロット	充足数 (%)	充足要素
A1: フィラー等	6 (6.4)	あ、あの／あ、すみません／あ、どうも／あの／あのう／えっと
A2: 呼びかけ	10 (10.6)	さん (3) / 君 (?) (2) / 閣下 / 先生 (ですか?) (2) / 先輩 / (の) みなさん
A3: あいさつ	14 (14.9)	はじめまして (6) / こんにちは (5) / こんばんは / よろしく / はいはい
A4: 応答	5 (5.3)	そうです (2) / どうぞ / はあ / はい
B: 1人称要素	22 (23.4)	私は (8) / 私、(5) / 僕、(3) / 私が (2) / 私の名前 (は) (2) / 名前は / 僕は
C1: 動詞要素	10 (10.6)	～した (8) / ～している (2)
C2: 名詞要素	20 (21.3)	所属先 (8) / 職種 (7) / 職位 (3) / 出身地 / 著作
D: 名前要素	94 (100)	姓 (40) / 姓名 (39) / 名 (15)

仮に、充足率10%以上のスロットを選び、かつ、それらを今回のデータで頻度が高かった要素で埋めると、たとえば、以下のような文が得られる（作例）。

- (14) みなさん、はじめまして。私は、今度こちらの支店に着任した ABC 商
事の佐藤（太郎）です。

D の名前要素について言うと、姓名の両方を言うことは必ずしも問題ではないが、その場合、やや大げさな名乗りの印象を与える可能性がある。頻度データでも姓名と姓のみはほぼ拮抗しているので、上の例では名のほうをカッコに入れている。なお、『みんなの日本語』の対話スキットでは、「こちらはマイク・ミラーさんです。」という他者による紹介を受けて、ミラー氏が「マイク・ミラーです。」と自己紹介しているが、こうした場合は姓のみにした方がより自然に聞こえるかもしれない。

また、B の部分に注目すると、係助詞の代わりに格助詞を使うことや、係

助詞を落とした言い方もかなり多いことがわかる。係助詞を落とすと、全体としてみてくださいのニュアンスになり、未充足のロットが増え、たとえば、下記のように文も短くなる（作例）。

(15) こんにちは、私、玲子です。

(14) のようなタイプか、(15) のようなタイプかは、自己紹介を行う人物の性別・年齢や社会的な属性、また、想定される聞き手との社会的関係性によって決定される。『みんなの日本語』のマイク・ミラー氏について言えば、アメリカ出身で日本の会社で働きながら日本語を学んでいるという設定が用意されているので、(14) のようなスタイルの自己紹介がより適切であろう。

Cに関して、前述の対話スキットでは、ミラー氏が、「初めまして。マイク・ミラーです。アメリカから来ました。どうぞよろしく。」と発言しており、A3だけでなくC1の内容も取り込まれている。C1要素については、おそらくは学習上の負担を下げるために独立した文にしたと思われるが、今回のコーパス分析結果をふまえると、C1を自己紹介文の中に埋め込むことで、より自然な自己紹介になった可能性もある。

日本語教材に限らず、外国語の教材にはリアルな文脈性の希薄な例文が多いとされ、「わたしはマイク・ミラーです」についても（少なくともこの個所だけを取り出せば）おそらくは同様の制約がある。もちろん、学習者のレベルを考えると、初級の第1課に置く例文として、これを大きく変えることは現実的ではないが、仮に、もう少し後の段階で、この文型を再度指導する機会があるならば、(14) や (15) のような例文を使い、日本語における自己紹介のより自然なあり方に触れさせながら、本稿で論じたような諸点について教師の側で追加の解説を加えるといったことも考えられるだろう。

小山（2008）は多くの日本語教科書が文法（文型）積み上げ方式になっていることを指摘したうえで、これからの日本語教育には、(1) 学習内容をその場で即座に習得させる方式から同じような内容を何度も繰り返させ段階的に習得を深めさせる方式へ、(2) 文法シラバスや機能シラバスから話題を核としたトピックシラバスへ、(3) 一回式の学びから「スパイラル・カーブ型」の学びへ、という発想の転換が必要であると述べている。

たとえば、自己紹介文についてもこうした発想に基づく新しい扱いが考えられるだろう。実際、Common European Framework of Reference for Languages

(CEFR) を「日本の言語使用場面に合わせて具体化した」国際交流基金の JF Can-do リスト (2016年 6 月版) では、自己紹介を含む活動は12種に上り、CEFR レベルも A2 から B2 まで広く分布している。下記はその一部である。

表8 自己紹介を含む can-do タスクの一例

No	CEFR	活動	カテゴリ	トピック	JF Can-do
22	A1	産出	経験や物語を語る	自分と家族	日本語の最初の授業など、初めて会う人のグループの中で自己紹介するとき、自分の家族の人数と構成を、簡単な言葉で言うことができる。
456	A2	産出	経験や物語を語る	自分と家族	初めて会った人の前で自己紹介するとき、自分や家族がどこに住んでいるか、何をしているかなど、短い簡単な言葉で話すことができる。
359	B1	やりとり	情報交換する	仕事と職業	取引先で、名刺を交換しながら、名前、所属、業務内容など、工作上必要な情報などについて、ある程度詳しく自己紹介し合うことができる。
195	B2	やりとり	社交的なやりとりをする	言語と文化	ホームステイ先で簡単な自己紹介をした後で、日本との出会い、日本語学習の経験、日本滞在中の予定など、さまざまな話題についての質問に、苦勞話や抱負を交えて答えることができる。

たとえば、こうした一連の「スパイラル・カーブ型」の学びの過程の中で、教師は、「わたしはマイク・ミラーです」という自己紹介文をその都度様々にアレンジし、そこに、豊かな言語的肉付けを与えていくことができる。コーパスはその際に大きな助けとなりうる。

4. まとめ

以上、本稿では、日本語コーパスの代表格である BCCWJ の概要を紹介した後、コーパスの教育応用の可能性を具体的に示すべく、初級の日本語教材に出現した「わたしはマイク・ミラーです」という文をサンプルとして基礎

的なコーパス調査を行い、関連する5つの観点について検証した。

その結果、RQ1（1人称単数代名詞としての「ワタシ」）については、各種の代名詞の中で、語彙素の「ワタシ」の頻度は最も高かったものの、占有比は6割程度にとどまり、「ボク」や「オレ」といった男性代名詞もあわせて4割近く出現していることが確認された。

RQ2（「ワタシ」の表記法）については、教科書に用いられた「わたし」の占有比は2割程度にとどまっており、漢字表記の「私」（73%）が最も典型的な表記形であることが示された。

RQ3（外国人姓名の表記法）については、姓と名を区切る補助記号として、ハイフン、ダブルハイフン、ナカグロの3種が存在するものの、実際には、ナカグロの占有比が97%に達しており、新聞や雑誌ではすべてがナカグロとなっていることから、教科書でも使われているナカグロが現代日本語における外国人名の標準的な表記補助記号になっていることがわかった。また、残りの2つの補助記号のうち、ダブルハイフンは専門的・学術的な文脈で特に歴史上の有名人などに言及する際に相対的に出現率が高まることがわかった。

RQ4（自己紹介文が出現するテキストタイプ）については、「<姓/名>です。」の文型は、国会会議録とブログにとりわけ多く出現するものの、その他の一般的テキストタイプでは出現が必ずしも多くないことが示された。

最後に、RQ5（「私は<姓名>です」文型の形態）については以下の典型的形態の存在が確認された。

[A：前部要素（A1：フィラー等／A2：呼びかけ／A3：あいさつ／A4：応答）] + [B：1人称要素] + [C：先行形容要素（C1：動詞要素／C2：名詞要素）] + [D：名前要素] + です。

また、このうち、BとC1、C2のスロットは全用例の20%以上で埋められること、また、A2およびA3のスロットは10%以上で埋められることがわかった。さらに、Dについては、姓名を共に明示する場合と姓のみを明示の場合がほぼ同等であることも明らかになった。

本稿で試みた検証はごく簡易なものであるが、一見平易で当たり前に見える文についても、コーパスを使うことで、様々な新しい言語事実が得られることを示している。もちろん、教材における例文の適否は、学習の段階や学

習者のニーズによって総合的に決定されるべきもので、コーパスから得られた事実をすべてそのまま例文や教材に反映すればよいということにはならない。しかし、経験を積んだ教師が教材に盛り込まれた言語材料を自ら吟味し、独自の視点で指導を深めていくうえで、コーパスが有力な補助具になることは確かであろう。実際、中俣（編）（2017）のように、「単にルールを教えるだけでなく、典型的な例を、根拠を持って示すことが必要」という立場から、大型コーパスの分析によって、各種の文法項目の典型構造を特定しようとする試みも始められている。今後、日本語コーパス研究がさらに進化し、日本語教育に今まで以上に有益な貢献をなすことを期待して本稿を閉じたい。

引用文献

- 石川慎一郎（2012）『ベーシックコーパス言語学』東京：ひつじ書房。
- 石川慎一郎（2017）『ベーシック応用言語学：L2の習得・処理・学習・教授・評価』東京：ひつじ書房。
- 荻野綱男（2017）「社会言語学」計量国語学会（編）『データで学ぶ日本語学入門』（pp. 56-68）東京：朝倉書店。
- 岸野英治（編）（2019）『ウィズダム和英辞典（第3版）』東京：三省堂。
- 国際交流基金（2016）「JF Can-do一覧表」
<https://jfstandard.jp/publicdata/ja/render.do#sec03>（2018年12月1日閲覧）。
- 小林敏（2015）「日本語組版とつきあうーその46：片仮名の人名表記の問題」
<https://www.jagat.or.jp/archives/10280>（2018年12月1日閲覧）。
- 小山悟（2008）「プロフィシアンシー重視の教材開発：トピック・ベースのシラバス・デザイン」鎌田修・嶋田和子・迫田久美子（編著）『真の日本語能力をめざして：プロフィシアンシーを育てる』（pp. 184-209）東京：凡人社。
- 中俣尚己（編）（2017）『コーパスから始まる例文作り』東京：くろしお出版。
- 李在鎬・石川慎一郎・砂川有里子（2018）『新・日本語教育のためのコーパス調査入門』東京：くろしお出版。

いしかわ しんいちろう
（神戸大学全学基盤系教育基盤域教授）