国立国語研究所学術情報リポジトリ

学習者コーパスの公開と公開後の運用

メタデータ	言語: Japanese 出版者: 研究社 公開日: 2026-07-17 キーワード: 学習者コーパス, コーパス構築, コーパス公開方法, 公開後の利用促進 作成者: 須賀, 和香子 メールアドレス: 所属: 総合研究大学院大学博士後期課程, 国立国語研究所リサーチアシスタント
URL	https://doi.org/10.15084/0002000639

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.



第15章｜学習者コーパスの公開と公開後の運用

本章は学習者コーパスを公開するにあたって留意すべきこと、また公開後にコーパス構築者が担うべき役割について述べるものである。コーパス公開にあたって重要なのは、利用者の利便性を考えることである。利便性を高めるために用意すべき資料やデータ整備、公開方法について示す。また、コーパスは公開して終わりではない。多くの人に使ってもらうコーパスにするために、公開後に行う広報活動や拡張などについて述べる。

KEYWORD
｜学習者コーパス｜コーパス構築｜コーパス公開方法｜公開後の利用促進｜

□1　はじめに

ここまでの章で、学習者コーパスの意義や役割そしてその歴史、また、国立国語研究所が現在開発している学習者コーパスの一つである『日本語学習者縦断作文コーパス（W-CoLeJa）』の設計や活用の方法などについて述べてきた。最後の章となる本章では、構築したコーパスを公開するにあたって留意すべきこと、そして、公開した後にコーパス構築者が担うべき役割について述べる。

国立国語研究所で現在（2025年7月時点）公開している日本語学習者コーパスは、迫田久美子氏をリーダーとする国立国語研究所基幹型共同研究「多文化共生社会における日本語教育研究」をはじめとするプロジェクトで構築した『多言語母語の日本語学習者横断コーパス（I-JAS）』と『中国語・韓国語母語の日本語学習者縦断発話コーパス（C-JAS）』、そして、国語研が北京外国語大学北京日本学研究センター、および北京師範大学外国語言文学学院と交流協定を結び、3機関共同で構築した『北京日本語学習者縦断コーパス（B-JAS）』の3つがある。W-CoLeJaは、これらに続くコーパスとして現在構築中であり、本章執筆時においてはまだ公開に至っていない。そのため本章では、国立国語研究所でプロジェクト非常勤研究員としてI-JASやB-JASのプロジェクトに属し、構築・公開に携わってきた筆者が、これまでの2つのコーパスを公開する際の経験や考え

てきたことをもとに、コーパス公開時・公開後に留意すべきことについて述べることとする。

□2　日本語学習者コーパス公開の意義

日本語学習者コーパス開発の意義については第1章で述べられている。学習者と接する多くの日本語教師は、学習者特有の誤用の傾向や習得の過程、特徴などがあることを、現場の中で肌身で感じている。日本語学習者コーパスには学習者の習得過程の日本語が収録されているため、これを利用することで、日本語教師がこれまで感覚的に感じてきたことを証明し、より深く研究していくことが可能となる。また、前川（2013）はコーパスから予想外の発見があると述べている。学習者コーパスにおいても例外ではなく、学習者の実際の日本語を見ることによって、これまでに気が付かなかった発見を得ることができる。これが学習者コーパス開発の意義である。

では、学習者コーパス公開にはどのような意義があるのだろうか。コーパスを公開するということは、そのコーパスを多くの人が利用できるということであり、学習者コーパスが多くの人に利用されることで第二言語習得についての研究の速度も深度も範囲も速く深く広くなっていくと思われる。

コーパスの構築には、費用に加え人的労力も時間もかかるものである（山崎ほか 2014）。多くの労力をかけて作り上げたコーパスを一部の人のみで利用するのではなく、公開することによって多くの研究者と共有し、コーパスに埋もれた多くの宝物を、さまざまな視点や手法によって発掘していくことができる。その発掘されたことを社会に蓄積していくことで、日本語教育に貢献することが可能となり、まさしくそれが学習者コーパス公開の意義であると筆者は考える。

次節からはコーパスを公開するにあたって留意するべきこと、そして、公開した後の役割について述べていく。

□3　コーパス公開前の留意点

前節で述べたように、コーパス公開の意義は、多くの人に使ってもらうことである。そのため、コーパスを公開する際には、まずはコーパス利用者の利便性を考えることが最も重要となる。利用者にとって使いやすいコーパスという視点から、公開のために準備すべきことと公開の方法について3.1〜3.4節で述べる。そのうえで利用者のみならず、データ提供者や構築者にとっても重要となる権利関係について3.5節で述べる。

3.1　調査やコーパスの概要説明書

コーパスを正しく利用し研究に活かすためには、コーパスの利用者は、そのコーパスがどのような目的で、どのような協力者からどのようなデータを収集したのか、また、どのようにデータを処理して構築されたものなのかを、正しく理解することが必要である。それを把握せずに研究を行うと、分析したデータの解釈を誤り、間違った結論を導き出してしまうことになる[1]。

また、コーパスを使った研究では、複数のコーパスを使用して分析し、総合的な研究を行う可能性も考えられる。例えば、一つのコーパスだけでは、データ収集時の制約や実施タスク、公開のためのデータ処理の方法など様々な要因で、研究に必要な論拠を十分に示せないことがある。その際、足りない論拠を補うために他のコーパスを使用し、複合的に分析するという方法も考えられる[2]。コーパス構築者は、そのような活用法のためにも各コーパスに関して十分な理解が得られる調査概要説明書やコーパス概要説明書を用意し、公開することが必須である。

正しいコーパス利用のため、また、これからのコーパス研究の発展の可能性を考えても、コーパスの詳細な情報が公開されていることは重要であり、さらには、利用者は各コーパスについて十分に理解して利用することが求められる。

3.2　公開する学習者の産出データ

学習者コーパスの主たるデータは、学習者が産出した発話データ、または作文データである。学習者コーパスは学習者が産出する日本語の実態がわかるコーパスだが、学習者が産出したそのままのデータを公開するのではない。公開にあたっては、研究データとして活用しやすくするための整備が必要となる。

必要な細かい整備は多々あり、それぞれコーパス構築の目的に合わせて方針を検討したうえで行わなければならない。例えば、データは研究時に該当部分を特定しやすくしたり、研究の照合や再現性に役立てたりできるよう文番号や発話番号を付与する。その際に重要なのが、文章の区切りである。このルール一つをとっても以下のようなことを考える必要がある。

1｜ 国立国語研究所では、2023年度のNINJALシンポジウムで、国語研で公開している各コーパスのそれぞれの設計を踏まえた正しい利用方法を「すべきこと」「してはいけないこと」の例とともに紹介するシンポジウムを開催した。その際の発表は動画として公開されている。
https://www.youtube.com/playlist?list=PLZfZgVvFbh1ZhKyCJzQM6k0XkdDKQq-tL

2｜ 迫田・佐々木・細井・須賀（2020）や、石川（2022）では、複数のコーパスを使用して、各コーパスの特徴を活かした多角的な分析がなされている。

W-CoLeJaのように、学習者の作文データのコーパスの場合、学習者が付与した句点を使用して1文を認定することができる。しかし、学習者の句点の付け忘れや、10行近くに及ぶような極端に長い1文を書いた場合など、個別の事例に合わせて、それらの処理方法について検討することになる。

発話データの場合は、より複雑な処理が必要となる。まずは発話の書き起こし作業が必須であり、そのためのルールを策定しなければならない。コーパス利用者のことを念頭に置き、活用間違いなどの誤用や言いよどみ、言い直しなど学習者の発話の特徴を残した書き起こしの方法に加え、作業者の負担も考慮してルールを策定する。書き起こしのルールは決まったものがあるわけではない。そのため、I-JASやB-JASでは、既存の発話コーパスの書き起こしルールを参考に、第二言語習得のどのような研究に適したコーパスにするのかを想定したうえで基本方針を立て、それに基づいて独自のルールを定めた[3]。

発話データの文の区切りや1文の認定についてもこのルール策定時点で検討することになる。音声には句読点がないため、どこを区切りとし、発話番号を付与していくのかを決めなければならない。I-JASやB-JASでは、対話は発話のターンを区切りとし、独話は1文であると判断できる箇所で区切るように規定をした。それに対して日本語母語話者の会話を収録した『日本語日常会話コーパス（CEJC）』では、より細かいルールを設け発話の区切りを規定している[4]。

文章データ、発話データともにデータ整備の際のルール策定は重要であり、基本方針のもと、出来うる限りの情報は維持できるよう書き起こしルールを策定していく。しかし、手を加えることにより失われてしまう情報も存在する。それを補うために、I-JASでは、作文データは文番号を付与したテキストデータに加えて、個人情報処理などの最低限の整備を行っただけのデータもPDFで公開し、データ整備によって失われてしまう改行位置や記号などの情報も研究に使用できるようにした。また、発話データについては、音声データも公開することで、声のトーンやイントネーション、ポーズの長さ、また、文字で表記するのが難しい曖昧な発音、相槌のタイミングや発話の重なりなどの、書き起こしに反映できなかった情報を使った研究も可能となるようにしている。

3| I-JASの書き起こしについては、I-JASウェブサイトの「関連データ」内にある「(2) 多言語母語の日本語学習者横断コーパス「I-JAS」 書き起こし規程集」に詳細を掲載
https://www2.ninjal.ac.jp/jll/lsaj/ijas-document.html
B-JASの書き起こしについては、B-JASのデータ配布サイトにある「B-JAS概要・タスク説明・文字化概要」に詳細を記載　https://www2.ninjal.ac.jp/jll/bjas/bjasindex.html

4| CEJC の転記ルールについては下記のサイトを参照。https://www2.ninjal.ac.jp/conversation/cejc/transcript.html

そして、もう一つ、公開データに形態論情報を付与するかどうかという点も、活用しやすいデータという視点での検討事項の一つとなる（第10章参照）。形態論情報とは、語の品詞や活用形、語種などの情報を指し、それらを付与しておくと、その情報を使って学習者が使用する語彙や表現の特徴、使用傾向、発達過程などを調べることが可能になる。形態論情報は文章データ・書き起こされた発話データ、どちらにも付与することが可能であるが、その際にも誤用や言い間違いなどにどのような情報を付与するのかを決める必要がある。また、形態論情報を利用した研究をする際に必要となる語数表[5]や語彙表[6]などを整備することも検討項目の一つとなる。

3.3　公開する関連データ

ここまで、コーパスの概要説明書について、また、学習者の発話や作文に関するデータの整備について述べたが、研究に必要となる情報はこれだけではない。他に、調査協力者である学習者に関連する情報も必要となる。

例えば、研究する際には学習者の母語、日本語レベル、学習歴などの属性の違いによる比較をすることがある。他にも学習環境、留学や来日経験、学習動機など様々な要因が日本語習得に影響することが考えられる。文章執筆時に使用したデバイスの情報や、執筆にかかった時間、使用したサイトやアプリなどの情報もデータ分析には役に立つ可能性がある。また、日本語レベルや学習環境などは、経年によって変化するものもあり、W-CoLeJaのような縦断調査ではこうした点にも留意し、複数回情報を収集することも考えなければならない。学習者の様々な情報が収集され、関連データとして示されていることは、コーパス利用者にとって大変有効なものである。

また、I-JAS・B-JASでは、J-CATとSPOTという2種類の日本語能力テストの点数の公開、背景情報、作文調査の際に行ったアンケートなど、様々な情報を公開している[7]。W-CoLeJaについてもSPOT[8]と、プロジェクトで開発した「かんたん日本語テスト」（第7章参照）を実施しており、それらの点数に加えて、背景情報なども公開する予定である。学習者コーパスが学習者という人間を重視するものである以上、学習者の産出データと、テスト結果やアンケートなどの背景情報をどのように組み合わせて分析するかが鍵になる。

5｜ 各データに出現する語の数を示した表
6｜ 各データにどのような語彙がどのぐらい出現するのかを示した表
7｜ I-JASの背景調査の詳細は迫田・石川・李（2020）に掲載されている
8｜ 実施が困難であった一部の国を除く

これらの関連データについては、データ収集時点の情報であることが重要であるため、どのような関連情報が必要かを事前に検討したうえで、データ収集と同時期に情報を収集しなければならない。データ公開の際には、それをどのような形で利用者に提示するかを考慮し、個人情報が含まれていないかの確認など、整備をしたうえで公開をすることになる。

3.4　データの公開方法

コーパスを公開するにあたっては、どこでどのように公開するかを考えなければならない。研究には再現性という観点が重要であり、研究の信頼性を保証するためにも、同じデータを使って同じ分析が可能である状況を保つ必要がある。そのため、コーパス構築者には公開したコーパスをなるべく長く維持することが求められる。このような「コーパスの永続性」も考慮に入れつつ、データの公開方法を検討しなければならない。現時点で考えられる公開方法としては下記があげられる。

3.4.1　インターネットによる公開

データをダウンロードできる状態にしてコーパスを公開する方法である。①コーパス公開のために独自のサイトを構築したり、コーパス構築者個人や所属するプロジェクトのサイトにデータを掲載したりする方法のほかに、②大学などの学術機関リポジトリに登録する方法、③特定非営利活動法人言語資源協会（GSK）などの第三者機関に登録を依頼し、公開する方法も考えられる。

① 独自に構築したサイトやプロジェクト関連サイトを通じた公開

独自に構築したサイトや、所属するプロジェクトのサイトでコーパスの公開をすることのメリットとしては、構築者自身が公開用画面を作成できるため、コーパスのイメージを利用者に伝えやすいことや、データの更新やサイトの改修なども自身で行えることなどが挙げられる。また、誰でも容易にダウンロードし利用できるという利用者側の利点もある。しかしその反面、ウェブサイトの維持管理が必要となるうえに、定期的なメンテナンスなども自身で行わなければならない。

② 機関リポジトリへの登録

学術機関リポジトリ（Institutional Repository）とは、大学および研究機関で生

産された電子的な知的生産物を捕捉し、保存し、原則的に無償で発信するためのインターネット上の保存書庫である（国立大学図書館協議会 2003）。構築したコーパスを研究成果として所属機関のリポジトリに登録することで長期間公開することが可能になるとともに、多くの人が無料で入手できるようになる。登録は所属する機関リポジトリへの登録方法に従って手続きを行うこととなる。場合によってはデータサイズなどの制約など条件がある場合がある。

③ 第三者機関を通じた公開

言語資源協会（GSK）は、言語資源の流通を促進することにより、言語資源を必要とする音声・自然言語処理分野の学術・研究・産業の発展、さらには、言語学分野の研究の推進に貢献することを目的として運営されている特定非営利活動法人であり、あらゆる「言語資源」を利用しやすい形で提供することを目指している（橋田・田中 2006）。このような、第三者機関を通じて公開する方法も考えられる。データ提供を希望する人は、GSKの場合はウェブページから申し込みを行い、協会と相談しながら公開方法を決定する。

【コーパス検索システムを介した公開】

3.2節で、形態論情報が付与されたデータは、その情報を利用して研究を行うことができると述べた。その形態論情報を利用してコーパス内のデータを検索・抽出することができるようにするためのツールを**コーパス検索システム**という。データの公開には、このシステムを経由して行うことも可能である。国立国語研究所では「中納言」という検索のためのwebアプリケーションを提供しており、I-JASやB-JASを含め現在（2025年7月時点）14のコーパスをこの検索アプリケーション上で公開している。



図1　webアプリケーション「中納言」

3.4.2　書籍での販売や配布で公開するもの

インターネットを介さずに公開する方法もある。『日本語教育のためのタスク別書き言葉コーパス（YNU 書き言葉コーパス）』（金澤 2014）は、コーパスについての概要や研究実例を紹介する書籍とともに、全作文データを付属するCD-ROMによって公開している。

また、データを物理的に配布するという方法で公開することも考えられる。物理的というのはCD-ROMや外付けのハードディスクに入れて配布する方法である。利用者の把握がしやすいという利点はあるが、利用者は手続きや登録の煩雑さがあり、また管理者側も手続きへの対応やデータ送付などの作業が必要となる。

これらの公開方法については、利用する際に登録を必要とするかどうか、無償か有償かなども検討しなければならない[9]。また、公開方法は、どれか一つの方法を選ばなければならないということではない。I-JASやB-JASについても、データをダウンロードする方法と、検索アプリケーション「中納言」を使った方法の2つの方法で公開している。どのような形で公開することが利用者にとって良いのか、データの管理はどこまでできるのかということを考えて、自身に合った公開方法を検討する必要がある。

また、各公開方法の検討に合わせて、公開するためのデータをどのようにまとめるのかも考える必要がある。例えば、ダウンロードする場合は、データを一つずつダウンロードするのか、データセットを作ってまとめてダウンロードするのか、また、検索システムの場合は、その表示画面をどのように見せるのか、どのような項目で絞り込みを行えるように設計するか、などを考えなければならない。これらはコーパス構築者のデータ利用に対する考えを反映するものであると筆者は考えている。

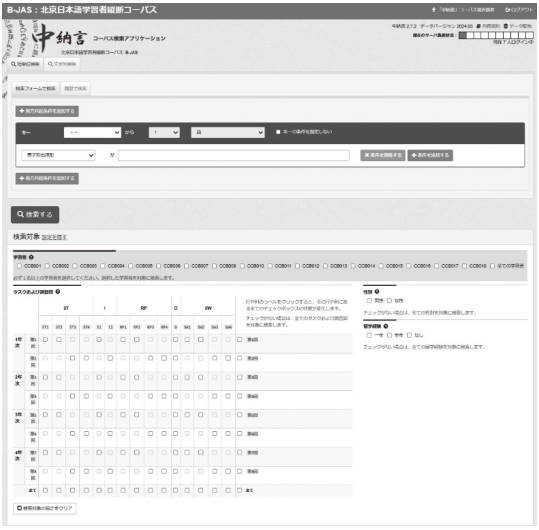

図2　「中納言」のB-JAS検索画面

例えば、B-JASのような縦断コーパスでは、ダウンロード用のデータをまとめる際、学習者ごとに第1回～第8回の調査データをまとめるのが良いのか、また

9｜ 有償のものでも実費相当の額を求めることが多い。

は、調査回ごとに17名の学習者のデータをまとめるのが良いのかということに対して迷い、検討を行った。その際、利用者の利便性も考えつつ、B-JAS構築者として、利用者にB-JASをどのように見てもらいたいかという点を考えた。縦断コーパスであるB-JASとしては、利用者には、ぜひ個々の学習者の成長に着目してほしいという思いがあり、それを伝えるためにも、学習者ごとのデータセットを作ることにした。そして、「中納言」のB-JASの検索画面も同様に、個々の学習者への着目を期待する意図から「学習者の選択」を絞り込み条件の最上部に設置した。

もとより、データをどのように使うのかは利用者の自由であり、それを制限するものではない。しかし、筆者は実際の構築経験から、どのようにコーパスを見せるのかという点には、構築者の思いや意図が含まれているということを付記しておきたい。

3.5　権利関係の処理

ここまでは、利用者にとって利活用しやすいコーパスを念頭に考えてきたが、それだけではなく①データ提供者の人権保護のための**個人情報の処理**や、②コーパスの**著作権の明示**、③コーパス利用の際の条件やルールを示した**利用規約の明示**についても考えておく必要がある。

①　個人情報の処理

1点目は個人情報の処理である。データを収集するにあたり、調査協力者には同意書へ署名をしてもらうが、その際には、おそらく「データ公開は個人が不利益を被らないよう、個人を特定できない形で行う」という調査者側からの条件を示しているはずである。そのため、調査協力者は、データ利用やデータ公開に対して同意済みであると思われる（W-CoLeJaにおける同意書の詳細は第4章）。コーパス構築者は、信頼してデータを提供してくれた協力者の人権を守るため、個人情報の処理には慎重を期す必要があり、この処理作業は大変重要な作業となる。一度公開したデータは、誰がどのように利用するかわからない。そのため、何が個人情報にあたるのかということを十分に検討し、漏れがないように複数の目で確認しつつ処理を行うべきである。

個人情報については、個人情報保護法によって個人の権利や利益が守られるよう定められているが、その法律をもとに、データを取り扱う際の注意点をまとめた研究者のためのハンドブック『オープンサイエンスのためのデータ管理

基盤ハンドブック』(国立情報学研究所 2022) が公開されている。これによると、個人情報とは、「一般人の判断力や理解力をもって、生存する具体的な人物と情報との間に同一性を認めるに至ることができる」ものと述べられている。それには、住所や電話番号、メールアドレス、ニックネームなどの具体的なものだけではなく、いくつかの情報を組み合わせることで特定の個人を識別することができるものも含まれる。そのため自由会話などのデータでは、全体的な内容にも注意を払い、個人情報箇所を特定していく必要がある。

I-JASやB-JASの個人情報についても、個人を特定できる情報は伏せた。人名だけではなく、所属大学名[10]や生年月日、また生年月日を連想させるようなイベント名なども伏せている。他に、出身地の地名については、その地域の人口や規模などを調べ、個人特定に至る可能性があるものは伏せるように配慮した。

伏せ字をするにあたっては、個人情報にあたる箇所を単に「個人情報」と表記するのではなく、上位語を用いて、「人名」「大学名」「日付」「イベント名」といった表記にした。そうすることで、伏せ字化されていたとしても、その箇所を含む文脈の内容がわかるように工夫した。各伏せ字には番号を付与し、同一の物事は同じ番号で示すことによって、文脈の理解を妨げないようにした。また、人の名前を呼ぶ際に付随している敬称は残すことで、敬称の研究などへの影響を避けるよう配慮した。

個人情報については、単語レベルだけではなく、文章の内容的に検討が必要になるものもある。データ収集の際には、会話の開始前に「言いたくないことは言わなくてもよい」と協力者に説明しているが、それでも収集したデータには、公開すべきではない内容や公開することで個人に不利益が生じる可能性がある内容 (過去の病歴や個人の思想など) が含まれていることがある[11]。そのような内容が出てきた際は、データ整備に関わる研究員間で相談をし、個人情報にあたる該当部分のデータをカットしたり、データそのものを公開対象から外したりする対応をした。

② 著作権の明示

2点目は、コーパス構築者の権利の明示についてである。まずは、データ収集時に調査協力者に対して著作権の所在を明らかにし、データの利用方法や公

10| I-JASでは調査を行った大学を公表していないため、所属大学を伏せたが、B-JASは、その限りではない。

11| 個人情報保護法では、「要配慮個人情報」として示されている。

開方法について調査協力者が同意していることが前提となる。そのうえで、コーパス構築者が保有する著作権について、利用者に向けて明示しておかなければならない。

著作権の保有者が、著作物の利用に対して、利用許諾の条件を明示する一つの方法として、**クリエイティブ・コモンズ・ライセンス**（CCライセンス）がある。これは、「インターネット時代のための新しい著作権ルールで、作品を公開する作者が『この条件を守れば私の作品を自由に使って構いません。』という意思表示をするためのツール」である（Creative Commons JAPANのホームページより）。

利用するための4種類の条件（表示：BY　非営利：NC　改変禁止：ND　継承：SA）が設定されており、権利保有者は、それらを組み合わせた6種類のライセンスの中から自身の希望するものを選択して表示することで、利用者に対して意思表示を行うことができる。利用者は権利者の意思表示を尊重し、ルールに基づいた利用をする必要がある。

③　利用規約の提示

3点目は利用規約の提示である。著作権の内容を含めてコーパス利用時の条件やルールを提示するものが、コーパスの利用規約である。これは利用者のためだけではなく、コーパス構築者の権利を守るためにも重要になるものである。コーパスを公開すると不特定多数の人がそのコーパスを利用することになる。それらの人に対して、権利の所在を明らかにし、コーパス利用に対するルールを示したり、トラブルなどがあった際の対応を利用規約の中で事前に明確にしたりしておくことで、コーパス構築者にとっても利用者にとっても、安心してデータを提供し、利用できるコーパスとなるだろう。

□4　公開後の責任

ここまでは、コーパスを公開するまでに準備し、考えるべきことを述べてきた。しかし、コーパスは公開して終わりではなく、利用者が使いやすいように、またトラブルがないように管理していくことも重要である。また、利用者もデータ提供者やコーパス構築者への敬意をもって利用すべきである。

コーパス公開後の構築者の役割や利用者の心構えについて、下記にまとめる。

4.1　コーパス構築者の役割

コーパス構築者の役割として、まず大切になるのは、利用者の声を拾えるよ

うな環境づくりを行うことである。利用者の声を受け取るためには、まず、問い合わせ先を明示しておく必要がある。この問い合わせ内容から、より多くの人に使いやすいコーパスへの改善のヒントを得ることができる。

問い合わせ窓口をどのように活かすかも含め、コーパス構築者の役割としては次のようなことが考えられる。

4.1.1 広報・講習会の開催

コーパス構築者の公開後の重要な仕事として、そのコーパスを多くの人に使ってもらうための広報活動があげられる。そのためには、まず、構築者が自らそのコーパスを使って研究を行い、発表などを通じてその存在を周知することが、有効な広報手段になると考えられる。

それに加えて、コーパスについての詳細な説明や使用法などに関する講習会の実施も重要である。I-JASやB-JASでも、また、国語研で公開している様々なコーパスにおいても定期的に講習会を開催している。

また、シンポジウムや研究会など大小の様々な企画やイベントを開催することも効果的である。これらの会を開催することで、利用者側は、コーパスについての理解を深めたり、構築者や研究者に直接相談をしたり、研究仲間とつながるきっかけができたりする。さらに、構築者にとっても、利用者の声を直接聞く機会になり、コーパスの弱点や問題点の把握、改善への工夫にもつなげられる。

4.1.2 データの修正や更新

コーパスは大量の言語データを扱うものであり、その大量の言語データに対して自動／人手で整備をし、情報を付与して公開をしている。データに不備がないよう注意しながら整備していても、中には公開後に不備が見つかることがある。不備が見つかった際にはそのデータを修正し、データを更新することを考えなければならない。データのメンテナンスを定期的に行い、不備を修正することで、利用者がより高い精度で分析できるコーパスとして更新されていく。

データ更新を行う際には、バージョンの管理が重要となる。公開直後からコーパスは分析に使われていくため、データにはバージョン情報を付け、データ更新のたびにバージョン情報も新しくすることを忘れてはならない。どのバージョンのデータを分析に用いたのかによって分析結果が変わってくるおそれがあるが、バージョン情報を付けておけば適切な追試が可能となり、そうし

た不安は解消される。また、更新を行った際には、更新箇所や更新内容を明らかにすることも大切な作業である。

コーパス利用者は、自身が研究に使ったデータがいつの時点のものなのかがわかるよう、分析の都度バージョン情報をメモしておき、データが更新された際には、自身が研究に使用したバージョンと、更新されたバージョンとでどこが異なるのかを把握し、場合によっては再分析を行うなどの対応を行う。

4.1.3 データの追加・データの整理など

コーパスを公開してしばらくたつと、先に述べた問い合わせ窓口に寄せられる質問などを通して、コーパスを実際に利用して研究する人からの意見が聞こえてくる。それらの声に耳を傾けることで、利用者が必要としていることが何なのかがわかり、ウェブページの見せ方に工夫をしたり、マニュアルを作成したりするなど、より使いやすいコーパスになるよう改善を行うことができる。例えばI-JASでは、「I-JASの全体の語数を教えてほしい」「形態論情報が付与されたデータは公開していないのか」「対話相手と学習者の関係性（初対面かどうかなど）について知りたい」など様々な問い合わせを受けた。そしてそれらの問い合わせにヒントを得て、語数表を作成して公開したり、形態論情報を付与した分類語彙表を公開したりして利用者の利便性を高めることができた。

また、I-JASでは、コーパス公開後に研究の補完をするデータも公開している。一つは、石川慎一郎氏により開発されたI-JAS for CIAというデータである[12]。これは、石川氏のゼミでI-JASを使って分析をする数名の学生がデータの処理に苦慮していたことから、石川氏と研究室のメンバーによってI-JASのデータを、中間言語対照分析（CIA）に使用しやすいよう再整備したデータである。

もう一つは、I-JAS-FOLASというデータである。これは、学習者のデータ分析を進める中で、誤用の原因が母文化にあるのかどうかの検証には新たなデータが必要であると考え、追加の調査を行って作成したものである[13]。これは、I-JASで調査を行った12言語のうちの5言語（中国語・英語・フランス語・韓国語・スペイン語）の母語話者同士に、I-JASと同じロールプレイとストーリーテリング、およびメール文のタスクを実施したものである。I-JASの学習者のデータに加えて、このI-JAS-FOLASのデータも分析することで、母文化の影響についてよ

12｜ I-JAS for CIAの詳細は石川（2024）を参照。

13｜「中納言」のコーパス選択画面にあるI-JASの備考欄に記載された「データ配布2」からデータの詳細の閲覧・ダウンロードが可能。（中納言アカウントが必要）

り深く考察することが可能となっている。

利用者の利便性を高めたり、研究の補助となるデータを追加で作成したりするなど、コーパスをより研究に活用しやすくするためのこうした拡張は、コーパス公開後の構築者の役割の一つであると考えらえる。

4.2　コーパス利用者の心構え

ここまで、コーパス利用者にコーパスを活用してもらうための構築者側の工夫について述べてきたが、最後に、コーパス利用者がコーパスを利用するときの心構えについて述べる。

多くのコーパス構築者は、そのコーパスの公開後に、コーパスの利用状況を把握する必要が出てくる。コーパス構築は科研費などの研究助成を受けて実施しているものが多く、研究助成に見合う実績を上げているかどうかを、研究助成の支援機関に報告する義務があるからである。事実、I-JASやB-JAS、そしてこれから公開するW-CoLeJaも、科研費や運営費交付金の補助を受けて構築しているものであり、利用者数や業績などの報告のために、適宜利用状況に関するデータを収集している。

これは、単なる報告義務というだけではない。開発したコーパスが多くの利用者によって研究活用されているということを成果として示すことで、コーパスそのもの、さらにはその構築自体にも良い評価がなされ、それが次の新しいコーパス構築に向けた新たな支援を得ることにつながるのである。そしてそれが、コーパス利用者の新たな研究への寄与となるだろう。次のコーパス開発に貢献するためにも、利用者は利用規約をよく読み、引用や報告等のルールに従って利用することをお願いしたい。

□5　おわりに

現在、コーパスは言語資源の一つとして考えられており、文化庁の審議会でも注目されるようになっている[14]。言語資源としてのコーパスという考え方は、その時代時代にどのような日本語が使われていたのかを記録するために、各時代を切り取って言語を収集し保存するという観点でコーパスをとらえたものである。本書・本章で述べている学習者コーパスのI-JASやB-JAS、W-CoLeJaは、ある課題に基づいて学習者が産出を行ったもので、完全に自発的なものではないため、言語資源としては、母語話者の産出した多くのコーパスと比べて異質

14｜ 文化庁国語審議会（2025）

な面があるかもしれない。しかし、現代では日本社会の中で、多くの定住外国人が、母語と日本語を使い分けながら、様々なストラテジーを用いて生活をしている。今後そのような非母語話者の多様な言語生活の実態が明らかになるような、多言語コーパスが作られる可能性もある。それを使って研究がなされる際には、母語別の学習者の日本語習得過程が調べられるI-JAS・B-JAS・W-CoLeJaのような学習者コーパスは大いに役立つであろうし、その意味でも学習者コーパスは重要なコーパスであると言える。

冒頭に、多数の利用者とコーパスを共有し、そこに埋まっている宝物をより多くの手で発掘し、発掘された発見が社会に蓄積していくことこそがコーパス公開の意義であると述べた。そうした目的を達成するためにも、学習者コーパス構築者として、公開するまでの十分な検討や必要な処理を踏まえたうえで公開を行い、さらに公開後にも、コーパスの在り方について考え、責任をもって管理していくことが求められる。

参考文献

石川慎一郎(2022)「時代変種と学習者変種の観点から考える日本語終助詞―時系列日本語小説コーパス「6121JFIC」と国際日本語学習者コーパス「I-JAS」を用いた統合分析の試み」『統計数理研究所共同研究リポート』456: 73–88.

石川慎一郎(2024)「中間言語対照分析(CIA)のためのI-JASダウンロードデータの加工―I-JAS for CIAの整備―」『言語資源ワークショップ発表論文集』1: 24–46.

金澤裕之(編)(2014)『日本語教育のためのタスク別書き言葉コーパス』ひつじ書房.

国立情報学研究所(2022)「オープンサイエンスのためのデータ管理基盤ハンドブック～学術研究者のための"個人情報"の取扱い方について～(第1版)」オープンサイエンスのためのデータ管理基盤ハンドブックにかかる検討会.
https://www.nii.ac.jp/service/handbook/ (2025.6.23閲覧)

国立大学図書館協議会 図書館高度情報化特別委員会ワーキンググループ(2003)『電子図書館の新たな潮流―情報発信者と利用者を結ぶ付加価値インターフェイス―』
https://www.janul.jp/ja/documents/report (2025.7.7閲覧)

迫田久美子・石川慎一郎・李在鎬(編)(2020)『日本語学習者コーパスI-JAS入門―研究・教育にどう使うか―』くろしお出版.

迫田久美子・佐々木藍子・細井陽子・須賀和香子(2020)「学習者コーパスを活用したモダリティ研究―日本語学習者の「かなと思う」の発達―」田窪行則・野田尚史(編)『データに基づく日本語のモダリティ研究』83–101, くろしお出版.

橋田浩一・田中穂積(2006)「世界の言語研究所(20) 特定非営利活動法人 言語資源協会(GSK)(日本)」『日本語科学』20: 107–111.

文化庁国語審議会(2025)「今後における日本語のデジタル言語資源の整備・活用の在り方」(報告)
https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/hokoku/ (2025.7.7閲覧)

前川喜久雄(2013)「コーパスの存在意義」前川喜久雄(編)『コーパス入門(講座日本語コー

パス　第1巻)』1–31，朝倉書店
山崎誠・前川喜久雄(2014)「コーパスの設計」前川喜久雄(監)・山崎誠(編)『書き言葉コーパス―設計と構築―(講座日本語コーパス　第2巻)』1–21，朝倉書店.

関連URL　(いずれも2025.7.7最終閲覧)

Creative Commons Japan
https://creativecommons.jp/
コーパス検索アプリケーション「中納言」
https://chunagon.ninjal.ac.jp/auth/login
多言語母語の日本語学習者横断コーパス(I-JAS)、中国語・韓国語母語の日本語学習者横断コーパス(C-JAS)
https://www2.ninjal.ac.jp/jll/lsaj/
特定非営利活動法人言語資源協会(GSK)
https://www.gsk.or.jp/
2023年度 NINJALシンポジウム「コーパス利用のべし・べからず」
https://www.ninjal.ac.jp/events_jp/20231105a/　(開催案内)
https://www.youtube.com/playlist?list=PLZfZgVvFbh1ZhKyCJzQM6k0XkdDKQq-tL
(発表動画)
日本語日常会話コーパス(CEJC)
https://www2.ninjal.ac.jp/conversation/cejc.html
B-JASデータ配布サイト
https://www2.ninjal.ac.jp/jll/bjas/bjasindex.html

(須賀 和香子)