

国立国語研究所学術情報リポジトリ

学習者コーパスの設計

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: 縦断作文コーパス, W-CoLeJa, 代表性, データ収集, データ公開 作成者: 石黒, 圭 メールアドレス: 所属: 国立国語研究所
URL	https://doi.org/10.15084/0002000627

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



第3章 | 学習者コーパスの設計

学習者コーパスをはじめ、学習者を対象に調査して収集した中間言語データを整理してデータベースを作成し、それに基づいて研究する場合、どのような手順を踏めばよいだろうか。本章では、『日本語学習者縦断作文コーパス (W-CoLeJa)』を開発した経緯を紹介し、学習者の言語データベース開発に必要な構想・収集・加工・公開の各工程に分け、作業の内容とそのさいの留意点について検討する。

KEYWORD

| 縦断作文コーパス | W-CoLeJa | 代表性 | データ収集 | データ公開 |

□ 1 学習者縦断コーパスを構想する

1.1 CoLeJa計画の背景

2000年代から2010年代の国立国語研究所（以下、国語研）は、コーパス研究の殿堂と形容できるほど、コーパス開発が活況を呈していた。日本語教育を専門とするスタッフもその例外ではなく、第3期（2016～2021年度）には、宇佐美まゆみ氏が『BTSJ1000人日本語自然会話コーパス』を、迫田久美子氏が『多言語母語の日本語学習者横断コーパス (I-IAS)』を、野田尚史氏が『日本語非母語話者の読解コーパス』と『日本語非母語話者の聴解コーパス』をそれぞれ開発しており（詳細は第2章参照）、学習者コーパス研究で大きな成果を上げていた。しかし、この3氏は第3期終了前後にいずれも定年退職を迎え、国語研内で学習者コーパスを担う人材が急に乏しくなった。

そこで、第4期（2022～2027年度）は、筆者が中心となって日本語学習者のコーパス開発を行い、第3期の研究を継続・発展させることにした。筆者自身はコーパス研究が専門ではないが、第1章執筆の迫田久美子氏のコーパス構築をお手伝いする過程で、氏の学習者コーパス開発の情熱に触れ、そのノウハウを見よう見まねで学ぼううちにその魅力にとりつかれ、自分でも学習者コーパスの開発に携わりたいと考えるようになった。12言語の母語話者を有し、世界20地域

で1,000名の学習者のデータを収集したI-JASの地理的横断性に太刀打ちすることは考えがたいが、日本語をゼロから学びはじめた学習者が日本語を習得し、大学4年間できわめて高いレベルに達する過程を粘り強く追うことならば自分にもできると考え、書き言葉の学習者コーパス『日本語学習者縦断作文コーパス』(以下、W-CoLeJaと呼ぶ)と、話し言葉の学習者コーパス『日本語学習者縦断談話コーパス』(以下、S-CoLeJaと呼ぶ)からなる『日本語学習者縦断コーパス』開発計画、いわゆるCoLeJa計画を始動した。

CoLeJa計画は、国語研の第4期機関拠点型基幹研究プロジェクト「多様な言語資源に基づく日本語非母語話者の言語運用の応用的研究」のもとで遂行された。W-CoLeJa(通称「ダブコレ」、WはWrittenの頭文字)の開発は、上記プロジェクトの「日本語学習者の作文の縦断コーパス研究」というサブプロジェクト、S-CoLeJa(通称「エスコレ」、SはSpokenの頭文字)の開発は、上記プロジェクトの「日本語学習者の談話の縦断コーパス研究」というサブプロジェクトで、それぞれ行われている。なお、CoLeJaという名称は、“Corpus of Learner Japanese”の略称であるが、耳で聞いたとき憶えやすい名称がよいと考え、「これじゃ!」という音になるCoLeJa(コレジャ)を採用した。

1.2 S-CoLeJaの概要

W-CoLeJaは本書全体で詳細に説明するので、ここではS-CoLeJaに先に触れておきたい。S-CoLeJaの前身は、迫田久美子氏が調査を企画・実施し、筆者が引き継いだ『北京日本語学習者縦断コーパス(B-JAS)』であり、S-CoLeJaはB-JASの拡張版である。B-JASは、『多言語母語の日本語学習者横断コーパス(I-JAS)』の縦断版であり、国立国語研究所、北京外国語大学北京日本学研究中心、北京師範大学外国語文学学院の3者の連携のもと、北京師範大学で2016年から2019年の4年間、17名の日本語学習者のデータを収集した。データ収集はI-JASの方法に準拠し、国語研の研究員が現地を訪問する対面形式で行われ、研究成果は林・徐・迫田主編(2024)にまとめられている。

17名の日本語学習者の良質な縦断データを、I-JASという大規模の横断データとの比較で分析できる環境を整えた意義は大きく、B-JASは第二言語習得研究に一定のインパクトをもたらすと思われる。しかし、中国の日本語学習者のみが対象であり、異なる母語話者同士の比較ができないという弱点がある。そこで、中国のほか、ベトナム、タイの計3大学で、2021年からデータの収集を開始した。当初は、B-JASと同様に年に2回現地に通り、対面の調査を4年間続

けることを想定していたが、COVID-19の感染拡大の影響で渡航が禁じられ、調査に着手できなくなった。そこで、オンラインで安定してデータを収集する方法を開発し、2025年に無事データ収集を完了した。中国、ベトナム、タイの大学の調査開始時の人数はそれぞれ31名、22名、34名、調査終了時の人数はそれぞれ26名、18名、17名である。一方、韓国の大学でも2024年からデータ収集に着手し、2027年にデータ収集が完了する予定である。調査開始時の人数は31名、2025年時点での人数は23名である。しかし、1年に2回、4年間続く調査の音声のデータ量が膨大で、文字化は困難を極めている。現状では、中国のデータのみ2027年度中に先行公開の予定であるが、研究チームとしても努力を重ね、可能なかぎり早期に公開したいと考えている。なお、S-CoLeJaは話し言葉のデータが中心だが、書き言葉のデータも並行して収集しており、両者の習得段階の比較ができるように工夫されている点も、これまでの学習者縦断コーパスに見られないS-CoLeJaならではの特徴である。

1.3 調査協力者の代表性

S-CoLeJaの紹介はここまでとし、ここからは、W-CoLeJaの構築を中心に筆を進めたい。W-CoLeJaの話にはなるが、独自設計のコーパスやデータベースを作りたいと考えている研究者にも参考になる内容であると思われる。

学習者コーパスを開発する場合、誰を対象に調査を行うかをまず考える必要がある。コーパス設計の場合、考慮すべきは**代表性**である。代表性とは、調査対象となる**母集団**を想定し、その母集団の実態が反映されるように**サンプリング**することである。たとえば、海外のJFL環境で学習する日本語学習者のコーパスを開発する場合、どのような国・地域でどのような学習者がどのぐらいの人数学んでいるのかを調べ、開発するコーパスがその縮図になるようにするのが理想である。

国際交流基金(2023)の調査によれば、日本語教育を実施している海外の国・地域数は141か国・地域に上り、海外の機関数は18,272機関、海外の日本語教師数は74,592名に、海外の日本語学習者数は3,794,714名に達している。学習者コーパスを開発するにあたっては、こうした海外の日本語学習者数を考慮し、それぞれの国・地域の学習者数を反映した構成になることが望ましい。国際交流基金(2023)に基づいて海外の日本語学習者数を国・地域別の順位に沿って示すと、次ページ表1のようになる。表1は上位10位に留まるが、上位10位だけで全体の89.5%を占めている。理想はこのような海外の日本語教育の現状を

踏まえ、上記の割合に合わせてデータを集めることである。そうすることで、世界の日本語教育機関で学ぶ日本語学習者の代表性を保証できる。

表1 海外の日本語学習者数(すべて)

順位	国・地域	人数	割合
1	中国	1,057,318	27.9
2	インドネシア	711,732	18.8
3	韓国	470,334	12.4
4	オーストラリア	415,348	10.9
5	タイ	183,957	4.8
6	ベトナム	169,582	4.5
7	米国	161,402	4.3
8	台湾	143,632	3.8
9	フィリピン	44,457	1.2
10	マレーシア	38,129	1.0

しかし、表1の数字の約半数は中等教育機関、すなわち中学・高校での日本語学習者である。本プロジェクトで対象としたいのは、書き言葉の第一言語習得がほぼ終了した成人日本語学習者の第二言語習得であるため、大学を中心とした海外の高等教育機関の日本語学習者数を基準とする必要がある。国際交流基金(2023)のデータによれば、表1とは順位が変わり、次の表2のとおりとなる。全体数は970,940名であり、上位10位で全体の89.2%を占めている。

表2 海外の日本語学習者数(高等教育機関のみ)

順位	国・地域	人数	割合
1	中国	557,153	57.4
2	台湾	61,990	6.4
3	米国	60,200	6.2
4	韓国	55,708	5.7
5	ベトナム	45,752	4.7
6	インドネシア	27,454	2.8
7	タイ	19,803	2.0
8	マレーシア	13,715	1.4
9	フランス	12,880	1.3
10	オーストラリア	11,301	1.2

今回の調査を代表性の観点から考えると、このバランスで調査をすることが理想であり、このバランスに近づけるべく、調査協力校の候補となる大学の先生方と交渉を始めた。その結果、過半数を占める中国から8大学、台湾と韓国から各2大学、ベトナムから3大学、タイから2大学、フランスは1大学の調査協力を得ることができた。それ以外では、英国、スロベニアからそれぞれ1大学の協力を得ている。

しかし、米国、インドネシア、マレーシア、オーストラリアは筆者が個人的なつながりを持つ大学がなかったために実施できなかった。こうした調査が個人のネットワークに左右されるのは残念なことであるが、縦断調査の場合、それは致し方ない面がある。1回きりの調査であれば、上記の国・地域の大学で友人のツテを頼り、無理を押してでも調査を行ったかもしれない。しかし、4年間の縦断調査を行う場合、現地の学生たちとつながりを4年間緊密に持続しなければならず、海外協力校の先生方にあいだに入ってもらい必要がある。しかも、その先生方に、調査協力者の学生たちと関係を保つために、4年間同じ学生の授業を持っていただくなど、かなりの負担を強いることになる。このような負担の重い役割をお願いすることは浅い人間関係では難しい。

したがって、縦断コーパスであるW-CoLeJaの場合、強い関係を構築できる先生方のいる海外協力校が第一条件で、そのうえで代表性という観点からバランスを考えて依頼を行っており、代表性という観点での限界は存在する。しかし、重要なことは、母集団の代表性を意識し、そこに近づける努力をすることである。今後、新たに学習者コーパスを開発される方も、代表性という観点から理想の姿を考えて、諸条件の制約のなかで理想に近づける努力をすることで、コーパスの質の向上にぜひつなげていただきたい。

1.4 調査協力校の選定

上述のように、学習者縦断コーパスを開発する場合、調査の過程でさまざまなトラブルが生じる可能性もあり、調査主体である国語研と海外協力校のあいだに強い信頼関係がなければ、4年間という長期にわたる調査を無事に完走することは不可能である。このため、まずは信頼関係を構築できる海外の調査協力校の教員への依頼を慎重に進めた。

負担の重い調査であるため、断られるケースも少なくなかったが、ご快諾くださる方のほうがはるかに勝った。結果的に、調査に協力してくださった先生方の特徴を確認したところ、

- ① 研究活動に前向きで、調査をきっかけに共同で研究をしたいと考える方
- ② 社会的責任から調査に協力したいと考える、要職にある責任感が強い方
- ③ 筆者と過去につながりがあり、感謝の気持ちでサポートを申し出られる方

のいずれかに当てはまった。調査協力校の先生方の調査負担は相当のものであったと推察されるが、上記①～③の理由から進んで協力してくださった方とは良好な関係が続き、途中で関係が悪くなったケースはなかった。世界各地で共同研究のパートナーに恵まれたのは、ほんとうにありがたいことであった。

もちろん、国語研のがわも調査継続のための多様な努力を重ねた。調査協力校の先生方には十分な説明を行い、こまめに連絡を取りつつ、先方の要望に可能なかぎり耳を傾けた。また、調査協力者の学生には、調査に支障のない範囲での作文へのフィードバック、ランゲージパートナーの斡旋、JLPT対策講座や学術・就職講演会の実施、調査参加証明書の発行など、費用ができるだけからない形での支援を行い、参加へのモチベーションの維持に努めた。具体的な取り組みについては第4章「調査協力者に調査を依頼する一協力校との連携と学習者のサポート」をご参照いただきたい。

1.5 世界の学習者コーパスの規模

コーパスに収録するデータを考える場合、調査協力者を何名に設定するかがポイントになる。調査開始当初は600名のデータを収集し、その3分の2の400名のデータを4年間取ることを目標としており、中国の8大学、台湾、韓国、ベトナム各2大学の計14大学での作文収集を考えていた。すでに紹介したB-JASにおいて、中国の1大学で作文調査を実施しており、そこで得たノウハウを生かしつつ、中国語圏の2大学で2020年に予備調査に着手した。その過程で調査方法を微調整したあと、残りの11大学で2021年から本調査を開始した。

しかし、調査を開始してわかったことは、学習者に調査を継続してもらうのが至難の業であるという事実であった。とくに、2021年と2022年におけるパンデミックによる教育現場の混乱の影響が大きく、調査に参加してくれている協力者の辞退がこの時期に急増し、2024年の卒業時点まで調査を継続できる協力者は3分の1程度になると、当初予想を下方修正せざるをえなかった。

このため、調査協力者の人数を確保する意味でも、コーパスデータの代表性を改善する意味でも海外の協力校を増やさざるをえないと考え、2022年から

2024年にかけて調査協力校を増やしていき、タイの2大学、ベトナム、フランス、英国、スロベニアの各1大学を新たに追加し、計20大学とした。その後、さいわいなことに、調査協力者の減少を食い止めることができ、2025年時点で、500名前後のデータが最終的に収集できると見込んでいる。この500名前後という数字は、海外の縦断作文コーパスと比較しても稀有の規模である。

日本語の学習者コーパスについては第2章ですでに詳述したので、英語の学習者コーパスを例にとると、たとえば、本章執筆時点での英語学習者のなかの世界最大規模の作文コーパスは、筆者が調べた範囲では、EFCAMDAT (EF-Cambridge Open Language Database) であった。このコーパスは、オンラインEFスクールに通う世界中の学生から提出されたエッセイで構成されており、第2版では174,743名の学習者から提出された1,180,310のテキストを収録している (<https://ef-lab.mml.cam.ac.uk/EFCAMDAT.html>)。

また、CLC (Cambridge Learner Corpus) は、Cambridge ESOLという英語試験の答案のエッセイを収集した大規模コーパスで、190ヶ国の130の異なる第一言語を持つ95,000名の受験生が執筆した作文を収録しているとされる (https://www.cambridge.org/elt/corpus/learner_corpus2.htm)。海外の学習者コーパスでは、このように試験の記述答案を活用して大規模化を図るものが多く、12,100本の母語別・レベル別にサンプリングされた作文を収録するTOEFL11 (<https://files.eric.ed.gov/fulltext/EJ1109982.pdf>) や、中国語能力試験HSK (漢語水平考試) を受験した受験者の11,569本の作文を収録する『HSK动态作文语料库』(hsk.blcu.edu.cn) も同種のコーパスである。ただし、上記のコーパスのうち、EFCAMDATは若干の縦断的性格は有するものの、厳密には縦断コーパスとは言えない。日本の中高生を対象にした英語学習者コーパスである『JEFLLコーパス (Japanese EFL Learner Corpus)』(<https://lingua-corpusnetwork.tufts.ac.jp/JEFLL2/>) や、アジアの英語学習者を広く対象としたICNALE (International Corpus Network of Asian Learners of English) (<https://language.sakura.ne.jp/icnale/>) はいずれも日本発で相当の規模を有するが、やはり厳密な意味での縦断コーパスではない。

一方、縦断コーパスは調査協力者の人数が相対的に少なく、比較的多いものでは、イタリアの多言語プロジェクトに参加した中学生を対象に、イタリア語、ドイツ語、英語の三つの言語での口頭能力・文章能力の3年間の発達を記録した縦断コーパスLONIDE (Longitudinal Learner Corpus in Italiano, Deutsch and English) (<https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/25>) や、日本の高校生を23ヶ月間、全8回の調査で追跡した英語スピーキング発達の経年的

記録である縦断コーパスLOCSE (Longitudinal Corpus of Spoken English) (<https://doi.org/10.69194/methodologysig.17.4>) があるが、前者は163名、後者は104名のコーパスである。上記二つのコーパスが100名を超える調査協力者を有する学習者縦断コーパスであり、開発中のものも含め、同等以上の規模のコーパスが存在する可能性もあるが、学習者縦断コーパスの収集は複数の機関にまたがって収集するのが難しいため、規模は相対的に小さくなる傾向があり、とくに複数機関を横断するコーパスはほとんどないと思われる。

第二言語習得研究は英語の習得を中心に世界各地で盛んであるが、その分析に不可欠な一定の規模を有する学習者縦断コーパスは、英語を含めてどの第二言語でも相対的に貧弱であり、レベル別コーパスを縦断コーパスと見なして研究が進められているのが現状である。しかし、それはいわば応急処置であり、学習者の第二言語習得過程を精密に観察する場合、相当数の同一学習者の同一課題を比較して行うのがもっとも精度の高い分析方法である。そうした定量的な分析を可能にするW-CoLeJaのようなコーパスが登場すれば、日本語という個別言語にとどまらず、世界の第二言語習得研究に一石を投じる可能性がある。

1.6 W-CoLeJaの規模

では、W-CoLeJaの調査規模はどのぐらいのものなのだろうか。概要は次ページの表3に示すが、あらかじめ表の見方について補足しておく。「大学ID」は、1文字目は母語、2文字目は調査地、3文字目は機関番号である。たとえば、CCAであれば、1文字目は母語のChinese、2文字目は調査地のChina、3文字目は機関番号のAとなる。機関番号はAから順に付されているが、I-JASの既存IDと重ならないように設計されている。調査協力者である学生個人のIDは、CCA001のように、大学IDのあとに数字3桁を組み合わせで示す形になっている。テーマについては後述するが、基本的にAテーマ、Bテーマの二つからなる。なお、Cテーマが一つだけあるが、これはB-JAS収集のさい、同時に収集した作文テーマである。「開始人数」は調査開始時の人数、「終了人数」は調査終了時の人数、「辞退者」は「開始人数」から「終了人数」を引いたものであり、()に入っているものは、調査が本章執筆時点で終了していない暫定数である。「開始時期」は調査開始時期であり、調査自体は基本的にその4年後に終了する。ただし、フランスとスロベニアの大学は3年制であるため、調査は3年間で終了する。なお、TTAとTTC、FFAとFFB、SSAとSSBは、調査開始時期は異なるが、同一の大学である。

表3 W-CoLeJaの調査規模

No	調査地	大学ID	テーマ	開始人数	終了人数	辞退者	開始時期
1	中国語簡体字圏	CCA	A	19	16	3	2020春
2	中国語簡体字圏	CCB	C	18	17	1	2016春
3	中国語簡体字圏	CCC	A	204	62	142	2021春
4	中国語簡体字圏	CCD	A	28	17	11	2021春
5	中国語簡体字圏	CCE	A	27	12	15	2021春
6	中国語簡体字圏	CCG	B	54	36	18	2021春
7	中国語簡体字圏	CCI	A	15	7	8	2021春
8	中国語簡体字圏	CCK	A	41	38	3	2021春
9	中国語繁体字圏	CTA	B	20	13	7	2020春
10	中国語繁体字圏	CTB	A	101	58	43	2021春
11	韓国語圏	KKA	A	59	21	38	2021秋
12	韓国語圏	KKB	B	14	5	9	2021秋
13	ベトナム語圏	VVA	A	27	24	3	2021春
14	ベトナム語圏	VVB	A	69	59	10	2022春
15	ベトナム語圏	VVC	B	97	68	29	2021春
16	タイ語圏	TTA	A	17	7	10	2022春
17	タイ語圏	TTC	A	21	(14)	(7)	2023春
18	タイ語圏	TTD	B	19	(12)	(7)	2024春
19	フランス語圏	FFA	A	36	25	11	2023春
20	フランス語圏	FFB	A	40	(22)	(18)	2024春
21	スロベニア語圏	SSA	B	16	2	14	2023春
22	スロベニア語圏	SSB	B	10	(5)	(5)	2024春
23	英語圏	EGA	A	2	(1)	(1)	2023春

調査開始時の人数を合計すると、954名になる。一方、現時点での調査終了時の人数は487名、辞退者は413名で、最終的な調査終了者は500名前後を予想している。さらに、ここにS-CoLeJaで収集した作文が合流するため、調査開始時の人数が1,000名規模の学習者縦断コーパスになる予定である。参加機関数も世界20機関、S-CoLeJa収集作文も加えると23機関になる。これは従来の世界の学習者縦断コーパスと比較しても、有数の規模である。

□2 データを収集する

2.1 作文のテーマ

どんなコーパスを開発するときでも調査設計は重要であり、コーパス開発の

目的によって**調査設計**は大きく変わってくる。W-CoLeJaの場合、世界各地の大学で学ぶ日本語学習者の第二言語習得過程、とくに文章執筆能力の発達過程が明らかにできるように、大学入学から卒業までの4年間、作文を経年的に収集するという目的があり、それに沿った調査設計になっている。

作文調査を考える場合、まず考えなければならないのが、作文のテーマである。W-CoLeJaの場合、第二言語習得過程を分析するコーパスであるため、異なる母語の比較、各学年での比較ができるように設計しなければならない。そう考えると、テーマは同じであるほうが望ましい。しかし、一つのテーマだけだと、分析結果がテーマに大きく左右され、客観性に不安が残ると考え、作文のテーマをAとBの2本立てとした。また、ジャンルについては、**体験文・説明文・意見文**に相当する3種類の課題をAテーマとBテーマそれぞれに設定し、それを4年間繰り返すように設計した。この三つのジャンルは文章力の基本を構成するものであり、この三つのジャンルに精通すれば実社会で通用する文章力が身につくと考えたからである(石黒 2022)。体験文は私生活でおもに必要となるジャンルであり、SNSやブログでの発信をはじめ、エッセイを執筆したりフィクションを創造したりと応用の範囲は広い。説明文はビジネスでおもに必要となるジャンルであり、ビジネス関係者に業務の報告をしたり顧客に商品の説明を行ったりするのに必須である。意見文は学術でおもに必要となるジャンルであり、論文やレポートなど、説得を目的に書く文章で求められる技術である。このように考えて、三つのジャンルを設定した。

それぞれのテーマは次ページの表4のとおりである。1年生でもそれなりに書け、4年生になっても飽きずに書け、インターネット等の情報をそのままコピーして済ませることは不可能な内容を考えた。安部 (2014) を参考にしつつ、調査協力者の「私」にとって身近な内容で、創意工夫を持って書けるような内容がよいと考えて、このような内容を設計した。

なお、CテーマはB-JAS用の特殊なものであり、「私」に関わる異なる内容を4年間で11回、書くように指示されている。また、Dテーマも存在する。これは、S-CoLeJa用の作文テーマであり、同時に収集する談話データと連動し、書き言葉と話し言葉の習得の違いを比較できるようにしたものである。

表4 作文のテーマ

	課題	文種	作文のテーマ：指示文	調査回
A	A1	体験文	うれしかったプレゼント：あなたがもらってうれしかったプレゼントは何ですか。教えてください。	01、04、07、10
	A2	説明文	私の好きな有名人：あなたが好きな有名人はどんな人ですか。その有名人を知らない人に紹介してください。	02、05、08、11
	A3	意見文	写真と動画：思い出を残すとき、写真と動画のどちらがよいと思いますか。理由も書いてください。	03、06、09、12
B	B1	体験文	思い出の旅行：あなたの思い出に残っている旅行はどんな旅行でしたか。教えてください。	01、04、07、10
	B2	説明文	行きつけの店：あなたがよく行く店はどんな店ですか。行ったことがない人に紹介してください。	02、05、08、11
	B3	意見文	過去と未来：もしあなたが、①5年前の自分に1時間会える、②5年後の自分に1分間会えるとしたら、①と②のどちらに会いに行きますか。理由も書いてください。	03、06、09、12

2.2 作文の文字数

文字数は各学年で次の表5のとおりである。「下限」は最低でもこの文字数以上は書いてほしいという文字数であり、習得状況から考えて何とか書けそうな文字数を設定した。ほとんどの作文はこの下限以上の文字数で書かれている。一方、「目安」はこの程度まで書いてほしいという文字数である。上限はとくに定めておらず、何文字書いてもよい。大学によっては既習者が入学してくることもあり、その場合は学年にかかわらず、3年生、4年生の文字数で書いてもらっている。また、比較のために日本語母語話者の作文データも収集しており、Aテーマが41本、Bテーマが24本であるが、上記と同様に3年生、4年生の文字数で執筆してもらっている。

表5 作文の文字数

学年	文字数の下限と目安
1年生	下限200字、目安400字、上限なし（何字でもOK）
2年生	下限400字、目安800字、上限なし（何字でもOK）
3年生	下限600字、目安1200字、上限なし（何字でもOK）
4年生	下限600字、目安1200字、上限なし（何字でもOK）

2.3 作文執筆の方法

作文収集では、手書きは採用しなかった。手書きで文章を書く機会は現代社会ではほとんどないと思われるからである。一方、現代社会で打ち言葉の入力に使われるスマホやタブレットも採用しなかった。まとまった長さの文章を書

くのスマホやタブレットは適していないと考えたからである。くわえて、過度に発達した予測変換機能が、学習者の文章力を把握するのに障害となるという理由もあり、まとまった長さの文章を書くのにもっとも一般的な、パソコンによるキーボード入力を採用した。一方で、キーボード入力を選択した積極的理由もあり、それはキーボードをどのように打ち、どのような手順でどのように入力したかのキーストロークをすべて記録するシステムを開発し、その解析をしたいと考えたからである。

その記録システムの詳しい内容は、第5章「入力インターフェイスの開発—EssayLoggerとMoodle—」をご参照いただくことにして、ここではその概要を簡単に紹介する。**EssayLogger**は、作文を執筆する過程でどの時間にどのキーを押したのかが記録できるようになっているアプリである。これを使うと、執筆した結果としてのプロダクトだけではなく、執筆する過程としてのプロセスも分析できる。すなわち、作文をテキストとしてだけでなく、ディスコースとしても分析でき、作文研究に新風を吹きこむことが期待される。また、多言語に対応しているため、それによって学習者が自らの母語による対訳を入力できるようになっている（第6章「母語対訳データの意義と活用法」も参照）。**EssayLogger**については、第一言語習得、第二言語習得を問わず、多くの作文コーパス開発者に無償で使っていただけるよう、近日一般公開の予定である。

なお、作文執筆のさいの条件として、学習者がふだん執筆している自然な環境を再現することに重点を置いている。執筆場所はとくに決めず、執筆時間の制限も設けていない。また、辞書使用に制限は設けておらず、執筆に必要な情報のインターネット検索も認めているが、例文をそのまま書き写したり例文集を使用したりすることは禁じている。**EssayLogger**自体も外部情報のコピー＆ペーストができない仕様となっている。また、**Google**翻訳や**DeepL**などの機械翻訳、**ChatGPT**や**Gemini**などの生成AIの使用も禁じているが、調査協力者の良心に委ねられているため、部分的に使用している可能性はゼロとは言えない。もちろん成績評価と関わらないことは周知しており、成績評価を上げるためにそうした支援ツールを使用するケースはまれであると思われるが、作文分析のさいは、外部情報が利用されている可能性がある点は留意されたい。

2.4 日本語能力の測定

学習者コーパスを開発するさいに、調査協力者の日本語能力の把握は重要である。どのぐらいの日本語能力を持っている調査協力者が作文を執筆したのか

がわからないと、作文の分析が困難だからである。

日本語能力を評価する方法には次の四つの方法が考えられる。

- ① 日本語の学習歴や大学の学年等を用いる
- ② JLPT (日本語能力試験) 等、外部試験の受験結果を用いる
- ③ TTBJ (筑波日本語テスト集) 等、簡便な試験の受験結果を用いる
- ④ データそのものから日本語能力を判定する

①は必要な情報ではあるが、これだけでは不十分である。個人で日本語を学んできた学習者の場合、総学習時間を特定するのはたいへん難しい。また、大学のような機関で学んできた場合でも、同じ学年でも日本語能力に差があることが少なくない。個人の能力と努力には差があるからである。したがって、①は当然のこととして、②～④と併用する必要がある。

②は情報としての精度は高いが、利用は難しい。JLPTの場合、年間の受験回数も限られ、受験している学生とそうでない学生がいるのがふつうである。一方、J-CAT (Japanese Computerized Adaptive Test) のような試験であれば、精度も高く、受験のさいの場所も時間も問わないのであるが、受験時間が長く、費用もかかる点がネックとなる。I-JASやB-JASでは採用されているが、W-CoLeJaやS-CoLeJaでは採用に踏みきれなかったのは、この2点が理由である。

③は情報としての精度は若干下がるが、費用がかからず、試験時間が短くて済むことからもっともよく用いられている。とくにTTBJに収録されているSPOT90の採用率が高く、W-CoLeJaやS-CoLeJaでも用いている。しかし、オンラインテストであるため、中国国内での受験が不安定であるという難点があった。このため、本プロジェクトでは、書き言葉の語彙能力と文法能力を測定することを目的とした「かんたん日本語テストごい」「かんたん日本語テストぶんぼう」という試験を自作し、SPOT90との対応関係がある程度見られるようにしてある。詳細については第7章「日本語能力測定試験の作成とその妥当性の検証―「かんたん日本語テスト」―」をご参照いただきたい。

④のように、データそのものから日本語能力を判定するという方法も気づかれにくいですが、考えられてよい方法である。たとえば、作文データがあれば、どのようなレベルの語彙・文法が使われているかを見ることで、執筆者の日本語力を推定することは理論上可能である。しかし、具体的に確立した方法があるわけではない。逆説的になるが、W-CoLeJaのような学習者縦断コーパスを用

いた精度の高い研究の蓄積が進むことで、作文から学習者の日本語レベルを正確に推定する方法が見いだせるようになると思われる。

2.5 背景情報の把握

学習者コーパスを開発するさいに、日本語能力の測定と並んで重要な情報が、学習者の**背景情報**である。詳しくは、第8章「学習者の背景情報の収集法」をご参照いただきたいが、W-CoLeJaでは3種類のシートを使って背景情報を収集している。

一つ目は「**フェイスシート**」、正式名称は「調査参加者フェイスシート」である。「フェイスシート」は、調査初回に1回だけ行った、学生の性別や出身など、基本的な個人情報を収録したシートである。

二つ目は「**情報シート**」、正式名称は「作文執筆情報シート」である。「情報シート」は、コンピュータの使用状況等を調べるために毎回、4年間で計12回の作文執筆時に実施している。PCを使って書く以上の機器の指定は行っていないため、どのようなPCや日本語入力システムを使ったのか、どのような辞書で調べ、どのようなWebページを検索したのかなどを記録してもらっている。

三つ目は「**アンケート**」、正式名称は「学習者アンケート」である。日本語での作文執筆にたいするビリーフや不安など、学習者の意識面を調べるために1年に1回、計4回実施した。日本語の作文能力に関する自己評価項目48項目への学習者の回答が収録されている。日本語学習歴(中学・高校・大学)や日本渡航歴(旅行・留学・居住)といった情報もこのシートに含まれている。

W-CoLeJaに限らず、学習者コーパスを開発する場合、この3種類のシートにあるような内容を記録しておく、分析のさいの参考になる。また、もし可能であれば、調査協力者に**フォローアップ・インタビュー**を行っておくと、コーパスの背景にある学習者の意識も知ることができ、有用である。

同時に、調査のサポートに当たる調査協力校の教員からの情報も分析のための貴重な情報となる。詳しくは第9章「学習者の学習環境と学習リソースを把握する」をご参照いただきたいが、たとえば、4年間のカリキュラムはどうなっているのか、どんな授業が開講されていて週何コマぐらい履修しているのか、何名ぐらいのクラス編成でどのような教授法で教わっているのか、作文の授業ではどんな教科書や教材を使っているのかなどが日本語習得に影響を及ぼしている可能性は高い。そうした情報は現場で教育に当たっている調査協力校の教員が熟知しているので、記憶が鮮明なうちに記録にまとめておいてもらうと分

析のさいに重宝する。

□3 データを加工・公開する

3.1 匿名化作業

コーパス開発は、データを収集して終わりではなく、データの収集の終わりが、データ公開にむけたデータ加工の始まりであると言える。データ収集後の初期作業は、話し言葉の場合は文字起こしであり、気が遠くなるほどの膨大な作業である。しかし、書き言葉、とくに今回のW-CoLeJaのようなPCによる入力データの場合、手書きの文字を打ちこむ必要はなく、入力データがそのまま分析データになり、作業自体はたいへん楽である。

ただし、作文の入力データをそのままインターネット等に公開するのは危険である。なぜなら、調査協力者の個人情報が作文データのなかに含まれる可能性があるからである。調査協力者自身、あるいは家族や友人の氏名など、個人情報の特定につながるような情報は伏せ字にしたうえで公開する必要がある。学会発表の予稿集や学会誌の投稿論文などで例文としてデータを使用する場合にも、個人情報が紛れこむことがないように、細心の注意を払う必要がある。

3.2 形態素解析

匿名化の次に必要になる作業は**形態素解析**である。形態素解析はすべての作文コーパスに必要な作業ではないが、形態素解析を済ませておくと、データの汎用性が飛躍的に高まる。

第一に考えられるのは検索性の向上である。文字列検索よりも語彙素で検索するほうが出現形の揺れなども吸収でき、誤検索や検索漏れが少なくなる。第二に考えられるのは分析の効率化である。総語数の把握にも便利であるし、品詞別の傾向を見ることができるようになる。文の構造を解析する句構造解析や係り受けの関係を解析する依存構造解析にも転用できるようになる。第三に考えられるのは教育への応用性である。語彙ごとくに例文を提示したり品詞別に練習問題を作ったりするのに便利である。国語研でコーパスを開発すると、コーパス検索Webアプリケーション「中納言」に収録されることが多く、必然的に形態素解析が求められることになるが、かりにそうした制約がない場合でも、こうしたメリットを享受するために形態素解析は検討されてよい。

日本語学習者の書いた作文の場合、学習者のおかしがちな文法などの誤用、辞書登録されていない固有名詞の使用、漢字の避用に伴う平仮名の多用などが

原因で、誤解析が起きやすいという問題がある。具体的にどのような問題があり、それにどう対処すべきかは、第10章「学習者作文における形態素解析の課題と対応」をご参照いただきたい。

3.3 誤用分析

誤用分析をコーパスで扱うのは難しい。どこまでを誤用とするのか、その判断が、誤用の定義によっても、認定者の違いによっても、学習者のレベルによって異なるからである。さらに、誤用とされたものをどう直すか、その直し方も認定者によって異なり、とくに初級学習者の作文の場合、執筆者に確かめないとわからないものも少なくないからである。W-CoLeJaに母語対訳がつけてあるのは、母語転移を分析するためだけでなく、学習者のもとの意図を確かめたいケースがあるという事情による。

第2章で見たように、誤用アノテーションが付されているコーパスもいくつか存在するが、そうした誤用アノテーションは唯一の正解ではなく、修正の一つの可能性と考えたほうがよい。もちろん、できるだけ安定的に誤用アノテーションを付す方法は検討すべきであり、第11章「日本語教師の誤用訂正はどこまで一致するのか」ではその試行錯誤の軌跡を紹介するが、私たちプロジェクトメンバーもいまだ暗中模索の状況にある。ただし、現代的には、安定的に誤用アノテーションを付す一つの可能性に生成AIの活用があり、第12章「生成AIはどのように学習者作文の誤用を訂正するのかー誤用分析のアノテーション付与のための一考察ー」ではそうした試みについて紹介している。

なお、形態素解析で見られる誤解析処理と、誤用分析で行う誤用アノテーションの付与は同じではない。誤用であっても誤解析を起こさなければ誤解析処理の対象にはならないからである。その意味で誤用アノテーションのほうが範囲が広く、その一部が誤解析処理の対象になると考えておくとよい。

3.4 印象評価

作文の質を考えると、避けて通れないのが読み手の評価である。誤用は学習者の作文の評価を下げる一つの要因にはなるが、誤用が多くても心に届く作文もある一方、誤用はなくても面白みに欠ける作文も存在する。同じ誤用でも、読み手の評価をさほど下げないものも、大きく下げてしまうものもある。

そこで、W-CoLeJaでは、作文そのものとは別に、一定の基準に基づいて読み手に評価してもらう評価データを作成した。それを用いることで、作文の表

現と評価の関係を見ることが可能になっている。詳しいことは第13章「印象評価情報の付与—クラウドソーシングの活用—」に譲るが、学習者の執筆した作文にたいし、7名の評価者に10段階の総合評価、5段階の理解容認度の評価をつけてもらうと同時に、内容（発想力、説得力、情報面での配慮、心理面での配慮、共感力の五つ）、構成（構成、一貫性、文のつながりの三つ）、表現（正確さ、ふさわしさ、豊かさの三つ）の三つの観点から計11の項目別評価について判定をしてもらっている。こうした印象評価データを収集することも、作文分析のさいに役に立つメタデータを作成するうえでの工夫の一つである。

3.5 コアデータ

ここで述べることは大規模コーパス限定にはなるが、たとえばW-CoLeJaで言えば、全部で8,000本前後ある作文データをすべて丁寧に読みこみ、分析するのは作業の労力から考えて現実的ではない。そこで、**定量的（量的）分析**を行うデータとは別に、**定性的（質的）分析**を行う限定的なデータセットを用意しておくことで、より安定した分析が可能になる。そこで、W-CoLeJaでは、全体のバランスを考えつつ、サンプルとなる**コアデータ**を選定している。

具体的には、中国語簡体字圏、中国語繁体字圏、韓国語圏、ベトナム語圏、タイ語圏、フランス語圏という六つの母語の学習者を選定し、各5名の学習者データを収録したcore_05と各10名の学習者データを収録したcore_10の2種類のデータを用意している。上述の誤用アノテーションや、印象評価アノテーションはcore_05にたいしてつける予定である。core_05と言っても、5名×年3回×4年間×6母語＝360本（実際にはフランス語圏は3年間なので345本）という本数になり、けっして少ない本数ではないが、相対的に見れば全体の1/20に満たない分量であり、定性的分析に適したデータセットである。

3.6 全文検索システム

全文データを用いるにせよ、コアデータを用いるにせよ、コーパス搭載のデータを駆使し、分析者が着目した表現の一覧を作成し、必要な例文にすぐに移動して分析することができれば、効率よく研究が進められる。第14章「W-CoLeJa活用環境の構築—全文検索システム『ひまわり』を用いて—」では、実際に収集した作文の分析ツール「ひまわり」の使用方法が紹介されている。W-CoLeJaも将来的には、山口昌也氏が開発したこの「ひまわり」を実装し、公開することで、分析者の便宜を図ることを目指している。

3.7 公開とメンテナンス

ここまで見てきた方法で構築されたコーパスは、**オープンデータ**が重視される**オープンサイエンス**全盛の時代、最終的にはWeb公開されることが理想である。しかし、Web公開するさいにはプライバシー保護はもちろん、データの公開形式、公開方法、メタデータ整備、バージョン情報付与など、留意すべき点もある一方、一旦公開されたデータは公開しっぱなしでは済まず、データ更新やサイトの管理、利用状況の把握やユーザー対応などが求められる。第15章「学習者コーパスの公開と公開後の運用」をご参照いただきたい。

□4 まとめ

本章ではW-CoLeJaの開発の経緯を概観し、構想・収集・加工・公開の各工程に分け、学習者コーパス開発のさいに必要な作業の内容とそのさいの留意点について紹介した。ここで述べたことは、W-CoLeJaのみならず、学習者コーパス一般の構築のさいに役立つ内容を多く含んでいると思われる。私たちのプロジェクトも失敗を重ねた結果、本章の見解に至っている。他山の石としていただき、よりよい学習者コーパスの開発に努めていただければさいわいである。

参考文献

- 安部達雄 (2014)「第11章 作文のテーマ選びの実際―常識を証明する活動を例に―」石黒圭 (編)『日本語教師のための実践・作文指導』104-113, くろしお出版。
- 石黒圭 (2022)「「物語る」「説明する」「説得する」3種類の文章を書く力、身につけるには 大事なのは読み手に伝える力」『朝日新聞EduA』.
<https://www.asahi.com/edua/article/14589946>
- 国際交流基金 (2023)『海外の日本語教育の現状 2021年度海外日本語教育機関調査より』.
<https://www.jpf.go.jp/j/project/japanese/survey/result/dl/survey2021/all.pdf>
- 林洪・徐一平・迫田久美子 (主編) (2024)『北京日语学习者历时语料库与日语教学 (北京日本語学習者縦断コーパス (B-JAS) と日本語の指導法)』中国国際广播出版社。

(石黒 圭)