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

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—

石川慎一郎（神戸大学）†

Resampling of the I-JAS Data for Contrastive Interlanguage Analysis: Compilation of "I-JAS for CIA"

Shin'ichiro Ishikawa (Kobe University)

要旨

「多言語母語の日本語学習者横断コーパス（I-JAS）」は、2020 年のリリース以降、日本語教育・第二言語習得の研究分野で広く使用されている。しかし、海外の学習者コーパス研究で広く実践されている計量的な中間言語対照分析（contrastive interlanguage analysis：CIA）は、I-JAS 研究ではあまり普及していないようである。この理由の一端は、I-JAS の習熟度データの複雑性と、I-JAS のダウンロード版テキストデータの扱いにくさにあると思われる。そこで、筆者は、習熟度を統制した CIA の実現のため、新しい習熟度指標で 1,000 人の学習者を再分類し、すべてのテキストデータを単一のエクセルシートに集約した「I-JAS for CIA」というデータシートを作成した。本稿は、「I-JAS for CIA」の構築過程とその利用法、また、研究応用の可能性について報告する。

Abstract

Since its release in 2020, the International Corpus of Japanese as a Second Language (I-JAS) has been widely used in research fields of Japanese language teaching and second language acquisition. However, quantitative contrastive interlanguage analysis (CIA), which is commonly practiced in learner corpus studies overseas, does not seem to be fully used in I-JAS research. Part of the reason for this may be due to the complexity of the learners' proficiency data offered in the I-JAS and the unwieldiness of the I-JAS downloadable text data. The author, therefore, has re-classified 1,000 learners based on a new integrative proficiency index and created a data sheet called "I-JAS for CIA," which consolidates all text data into a single Excel sheet to enable proficiency-controlled CIA. This paper reports on the process of constructing "I-JAS for CIA," its usage, and potential research applications.

１．はじめに

2020 年のリリース以降、「多言語母語の日本語学習者横断コーパス（I-JAS）」（迫田他，2016；迫田他，2020）は、多様な日本語学習者の L2 習得を科学的に議論する際の不可欠な一次資料として広く研究に応用されている。また、中国において縦断データを集めた B-JAS や、日本人の小中高大生の作文を集めた JASWRIC など、I-JAS と同一のタスクを用いて異なる種類のデータを集める動きも広がっている。一方で、筆者の知る限り、I-JAS を用いた研究の大半は、国立国語研究所の運営する統合的コーパス検索サイトである「中納言」上での用例検索や、ごく小規模な量的比較に留まっており、母語背景を異にする学習者間の相互比較が可能になるよう高度に統制的なデザインで構築されている I-JAS の真の価値は、いま

† iskwshin@gmail.com

だ十分に引き出されていないように思える。

海外の学習者コーパス研究では、中間言語対照分析（contrastive interlanguage analysis : CIA）と呼ばれるデータ分析法が広く実践されている。CIA には旧版と改訂版があるが、旧版にあたる CIA1（Granger，1996）では、NL（native language）と IL（interlanguage）、つまり、母語話者[1]による L1 産出と学習者による中間言語産出の比較、および、IL と IL、つまり、母語を異にする学習者同士の比較を行うことが提唱された。たとえば、中国語母語の日本語学習者を研究対象とする場合、日本語母語話者と比較するとともに、韓国語・英語・インドネシア語といった他の母語背景を持つ学習者と並行的に比較するのである。これにより、中国語母語の日本語学習者に固有の特徴と、広く学習者に共通する一般特徴の切り分けが可能になる。また、統計を援用し、2 群間での語彙頻度の差を悉皆的に調査することで、明確な誤用（misuse）だけでなく、特定の学習者群による過剰使用（overuse）や過少使用（underuse）が網羅的に抽出できる。大半のコンコーダンサに搭載されている特徴語分析（keyword analysis）機能を使えばこうした分析が容易に実行できることから、CIA は学習者の中間言語運用を俯瞰的に把握するための標準的な分析手法として広く普及するに至った。

その後、CIA1 は、比較の基準を母語話者に限定していた点が一部で批判されるようになり、2015 年に改訂された CIA2（Granger，2015）では、NL に代えて RLV（reference language varieties）という新しい概念が導入され、さまざまな RLV 同士の比較、さまざまな ILV（interlanguage varieties）同士の比較、さらにはこれら両者の比較を行うよう方向性が微修正されたが、根本的な構造は CIA1 から変わっていない。

このように、CIA の枠組みでは、母語話者（またはそれに代わる何らかの参照基準）との比較だけでなく、学習者間での比較も重要である。その際、多元的な比較の妥当性を担保するには、タスクや習熟度をそろえておく必要がある。この意味で、12 種の母語背景を持つ 1,000 名の学習者に同一のタスクを行わせ、その全員から、習熟度データと L2 産出データを集めている I-JAS は、まさに、CIA のための理想的な言語資源に見える。ただ、事態はそう単純ではない。

現在、I-JAS を利用するには、国立国語研究所が運営するコーパス検索アプリケーション「中納言」上で検索する方法と、テキストデータをダウンロードして自身で検索を行う方法があるわけだが、前者は、習熟度を統制したデータ比較に対応しておらず、後者には、(1)習熟度や母語による分類がなされていない、(2)フォルダ構造が複雑でデータが取り出しにくい、(3)テキストファイルの中に行コードと産出テキストが混在しており、ファイルのまま解析が行いにくい、といったいくつかの制約がある。結果として、一般ユーザーから見ると、I-JAS を CIA の枠組みで使うハードルはきわめて高いものになっている。

そこで、筆者は、これらの制約を可能な限り排除し、I-JAS を一般的な学習者コーパス研究、とくに CIA 研究の目的で使用できるよう、I-JAS ダウンロード版テキストデータの整理と統合を行い、I-JAS for Contrastive Interlanguage Analysis（I-JAS for CIA）という新しいデータシートを作成した。なお、I-JAS for CIA は、筆者が独自に作成したもので、I-JAS の開発者グループによる公式のリリース物ではないことをあらかじめ断っておく。

以下、2 章では、CIA において、学習者の習熟度を統制する必要があることを指摘した後、

[1] 近年の言語研究において、母語話者（native speakers）という概念の安易な利用に対して厳しい批判が加えられているのは周知の通りである（Chen et al., 2021 ほか）。本論文で母語話者という用語を使用する場合は「いわゆる母語話者」を指すものと解されたい。

I-JAS に含まれる J-CAT と SPOT90 の 2 種のスコアデータに関して、スコアの主成分化による合成（2.3 節）と、JLPT レベル推定値の合成（2.4 節）という 2 つの方法を示し、後者の利点を主張する。その後、後者の方法による学習者レベル分布を示す（2.5 節）。

3 章では、I-JAS のダウンロードデータの構造を示した後、サブフォルダからテキストファイルを一括で取り出して集約し（3.1 節）、単一ファイルにマージした後、それをエクセルシートに転記し（3.2 節）、行コードに含まれる属性情報を取り出すとともに（3.3 節）、母語と習熟度の情報を新規に付与し（3.4 節、3.5 節）、テキストの加筆部を削除するクリーニングを行い（3.6 節）、関連情報を追加して（3.7 節）、I-JAS for CIA データシートとして整備するまでの過程を示す。

4 章では、はじめに、作成された I-JAS for CIA の活用法を示した後（4.1 節）、研究実例として、I-JAS for CIA から抽出されたテキストを用い、習熟度を統制した上で、動詞頻度を手掛かりとして学習者の母語別分類を行った結果を示す（4.2 節）。

最後に、5 章において、全体の議論を整理する。

２．CIA と習熟度

２.１　習熟度統制の必要性

前述したように、CIA では、母語話者（あるいはそれに代わる参照基準）と学習者の比較に加え、母語を異にする学習者間の比較が重視される。どちらの比較でも重要になるのは、学習者群間で習熟度を均等にしておくことである。たとえば、母語話者と学習者の差を見る場合、X 言語話者群が上級者中心で、Y 言語話者群が初級者中心であれば、得られた差が母語によるものか、習熟度差によるものか判断できない。このことは、X 言語話者と Y 言語話者を比較する際にもあてはまる。

この点に関して、I-JAS のすぐれた特徴の一つは、日本語母語話者を除く 1,000 名の参加者全員が、J-CAT（項目応答理論に基づくテスト。聴解・語彙・文法・読解の 4 セクション各 100 点、合計 400 点）と、SPOT90（読み上げを聞きながら文中の空欄に入る平仮名を素早く選ぶテスト。合計 90 点）という 2 種の習熟度テストを受験していることである。内外の主要な学習者コーパスを見渡しても、学習者全員に本格的な習熟度テストを二重に受けさせている例はほとんどなく、これは I-JAS の特筆すべき利点と言える。

しかしながら、この優れた特徴が、逆に、I-JAS のデータを習熟度別に比較することを難しくしている面もある。実際、2 つのテストは、構成概念・内容・配点などがまったく異なっており、テストによって異なったレベル付けがなされている場合もあることから（一方では上級に、他方では中級に分類されるなど。迫田，2022，p.86 参照）、合算や平均化といった単純なスコアの統合もできない。そのため、せっかく価値ある習熟度データがありながら、それらを活用しない研究が多数を占めることになっているのではないかと推察される。

２.２　2 種のテストデータの性質

この問題を解決するには、何らかの方法で 2 種のテストスコアを合成する必要がある。それによって、学習者 1,000 名を何らかの習熟度バンドに分類することができれば、習熟度を統制した比較が可能になる。だが、まずはその前提として、2 種のテストがおよそ同等の能力を測っていることを確認しておくべきであろう。両者が根本的に異なる能力を測っているとすれば、スコアを合成すること自体が不適切な処理となるからである。

そこで、1,000 人のデータを対象に、J-CAT の 5 種のテストスコア、および、SPOT90 スコアの基本統計量を見たところ、以下の結果となった。

表 1　習熟度テストスコアの基本統計量

	J-CAT					SPOT90
変　数	聴解	語彙	文法	読解	総合	
サンプル	1,000	1,000	1,000	1,000	1,000	1,000
平　均	54.41	53.42	47.90	46.22	201.95	66.23
標準偏差	17.24	17.62	17.27	14.26	56.88[2]	12.06
変動係数	0.317	0.330	0.360	0.309	0.282	0.182

注目すべきは、J-CAT の平均点が満点に対して 47%～54%であるのに対し、SPOT90 の平均点が 70%以上になっていることである。テストとして見れば、J-CAT のほうが難度が高く（高得点がとりにくく）、SPOT90 のほうが難度が低い（高得点がとりやすい）。このため、変動係数（標準偏差を平均で除した値）において、J-CAT が 0.28～0.36 であるのに対し、SPOT90 は 0.18 程度でばらつきの幅が狭くなっている。

こうした違いはあるものの、以下の相関表が示すように、6 種のスコアはおよそ似た関係にある。

表 2　テストスコア間の相関

		J-CAT					SPOT90
		聴解	語彙	文法	読解	総合	
J-CAT	聴解	1.00					
	語彙	0.65	1.00				
	文法	0.63	0.72	1.00			
	読解	0.64	0.62	0.59	1.00		
	総合	0.86	0.88	0.86	0.82	1.00	
SPOT90		0.74	0.74	0.69	0.66	0.83	1.00

J-CAT 総合スコアと各セクションスコアの相関は 0.82～0.88、また、J-CAT 総合スコアと SPOT90 の相関は 0.83 となっており、構成概念や形式が異なっても、これらのスコアはおよそ似通ったものとなっている。これより、J-CAT 総合と SPOT90 は総体として類似した能力を測っており、2 種のスコアを合成することには一定の合理性があると判断できる。

次に問題になるのは、では、J-CAT スコアと SPOT90 スコアをどのように合成するか、という点である。筆者はまず、J-CAT の各セクションスコアと SPOT90 スコアを主成分化する方法を試みた。その後、J-CAT 総合スコアと SPOT90 スコアをそれぞれ日本語能力試験（JLPT）の推定レベル値に換算し、両レベル値の平均に基づいて、全体を区分する方法を試みた。最終的には、後者の手法が妥当性が高いと判断した。以下では各々の手法の概要を示したい。

２.３　2 種のテストデータの合成（1）スコアの主成分化による合成

前節の知見をふまえると、J-CAT のスコアと SPOT90 スコアの合成は、理論上、不可能ではないということになる。単位の異なる多変量データを一つにまとめる際には、しばしば、主成分分析の手法が使われる。そこで、線形結合している J-CAT 総合スコアを除き、J-CAT の 4 つのセクションスコアと SPOT90 スコアの全 5 種に対して主成分分析を実施したところ、全体の「総合力」と解釈される第 1 主成分（PCA1）が取り出され、寄与率は 73.6%で、

[2] 迫田他（2020）には J-CAT の標準偏差が 58.89 とあるが（p.82）、2024 年 7 月現在の公開データで検証したところ、56.88 となった。

5 変数の分散の 7 割以上を集約できることがわかった。

表 3　第 1 主成分負荷量

変　数	PCA1
J-CAT 聴解	0.853
J-CAT 語彙	0.874
J-CAT 文法	0.847
J-CAT 読解	0.815
SPOT90	0.897

すでに見たように、相互の相関が高いことから、5 変数の主成分負荷量もほぼ同等となった(あえて言うと、第 1 主成分に相対的に最も強く寄与するのは SPOT90 スコアであった)。

これをふまえ、1,000 人の各々について第 1 主成分得点を計算した。以下は中国語母語の CCH の 50 名の PCA1 値である。

表 4 個人学習者別の第 1 主成分得点（CCH のみ記載）

ID	PCA1	ID	PCA1	ID	PCA1	ID	PCA1	ID	PCA1
CCH02	1.05	CCH15	1.15	CCH25	−1.18	CCH35	2.12	CCH48	0.04
CCH03	1.59	CCH16	2.99	CCH26	−0.31	CCH36	−0.47	CCH49	−1.39
CCH06	1.47	CCH17	1.74	CCH27	0.18	CCH37	1.67	CCH50	−0.13
CCH07	2.86	CCH18	0.48	CCH28	−0.79	CCH38	1.27	CCH51	0.43
CCH08	1.84	CCH19	0.58	CCH29	−0.17	CCH39	1.29	CCH52	0.12
CCH09	3.14	CCH20	0.15	CCH30	−0.62	CCH40	0.60	CCH54	0.71
CCH10	0.10	CCH21	−1.22	CCH31	−0.38	CCH42	1.41	CCH55	2.84
CCH11	2.13	CCH22	0.81	CCH32	1.01	CCH43	1.06	CCH56	0.98
CCH12	2.66	CCH23	0.77	CCH33	2.16	CCH45	2.17	CCH59	1.54
CCH13	1.11	CCH24	−0.64	CCH34	0.52	CCH46	2.21	CCH63	−0.54

主成分化を行うことで、単一尺度で学習者の習熟度が比較できるようになった。ただ、主成分得点は平均が 0 になるよう調整された値に過ぎず、値自体に意味はないことから、主成分得点をもとに合理的な習熟度区分を行うことは困難であるという結論に至った。

２.４　2 種のテストデータの合成（2）JLPT レベル推定値の合成

筆者自身も、過去の研究において上述の主成分得点を用いたことがあるが、能力レベルを直接に示していない（たとえば、主成分得点の 1 や 2 が具体的にどの程度のレベルを表しているのかはわからない）点で、こうした値の活用範囲は限られる。実際、最近の I-JAS 研究においても、主成分得点の使用はまったく広がっていない。

この点をふまえると、数値自体に意味があり、数値を見れば、習熟度の具体的イメージが持てるような新たな値が必要ということになる。そこで、今般の I-JAS for CIA の開発では、両方のテストに示されている JLPT レベルとの換算表を参考に、新しい習熟度合成値を作ることとした。

下記は、「日本語テストシステム J-CAT」のウェブサイト内にある「J-CAT のスコアについて」（www.j-cat2.org/html/ja/pages/interpret.html）、および、「筑波日本語テスト集（TTBJ）」のウェブサイト内にある「得点の解釈」（ttbj.cegloc.tsukuba.ac.jp/p1.html）に記載されている対照表である（2024 年 7 月 30 日閲覧）。

J-CAT	JLPT 日本語能力試験
150-	N4
200-	N3
250-	N2
300-	N1

図 1　J-CAT 総合スコアと JLPT レベルの対照表

合計点	能力判定**	説明	日本語能力試験**
0～30	入門	日本語を学習したことがほとんどない。	なし
31～55	初級	ゆっくりであれば日常生活の基本的な日本語を理解できる。	N4，N5
56～80	中級	自然な話速度で日常的な場面の日本語がある程度理解できる。	N3，N2
81～90	上級	自然な話速度で幅広い場面の日本語が理解できる。	N1

図 2　SPOT90 と JLPT レベルの対照表

これらは大まかな対応に過ぎず、SPOT90 についても、JLPT レベルとの対照は「合格を約束するものでは」なく、「目安として理解」されるべきものと注記されているが、テスト開発機関が公式に出している換算表という点で、これらの値には一定の信頼が置ける。つまり、2 種のテストの点を、JLPT のレベル値に換算することで、性質の異なる 2 つのスコアを合算することができるのではないかと考えられるのである。

もっとも、上記の表では、JLPT の 1 段階に相当するスコアの範囲が広すぎ、かつ、J-CAT の N5 下限値や、SPOT90 の N2・N4 下限値などがはっきりしない。そこで、記載されているレベル値とスコアを手掛かりに両者の線形的な関係を推定すると、以下のような関係が得られる。

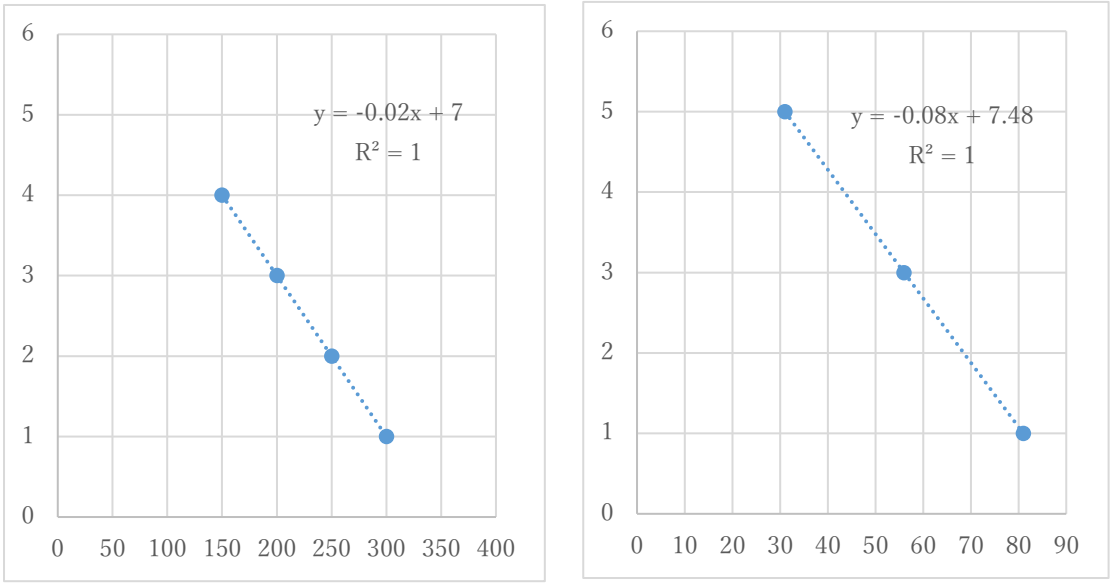

図 3 J-CAT（左）・SPOT90（右）のスコアと JLPT 換算レベルの関係

得られた回帰式を J-CAT と SPOT90 のテストスコアに適用すれば、学習者個々人につき、単純な 5 段階ではなく、連続指標化された JLPT 換算レベル値を 2 種取得することが

できる。これらを平均することで、図 1・図 2 の基準をそのまま当てはめた場合以上に、個々人の JLPT 換算レベルを正確に推定することができると考えられる。

下記は、CCH のサンプル学習者を対象として、図 1-図 2 に基づくレベル値の平均化（平均化手法 1）と、本研究で提案する回帰式を使ったレベル値の平均化（平均化手法 2 の結果を比較したものである。なお、手法 1 における SPOT スコアからのレベル値推定に関しては、56～80 点が N3～N2 とされているだけで、両者の境界は示されていないことから、56～80 の学習者には一律に中央値として 2.5 のレベル値を割り振る。また、いずれの手法においても、平均で得られたレベル値を最終的に段階レベルに戻す際には、詳細な差を反映できるよう、小数点第 1 位に四捨五入を適用する。たとえば、レベル値が 0.50～1.49 であれば N1、1.50～2.49 であれば N2、2.50～3.49 であれば N3 とみなす。また、実際の JLPT には設定がないが、便宜上、レベル値が 0.0～0.49 であれば N0（N1 を上回る母語話者レベル）、5.50～6.49 であれば N6（N5 に満たないレベル）とみなす。

表 5　2 種の平均化手法による J-CAT および SPOT90 テストスコアの合成

ID	スコア		JLPT レベル（手法 1）				JLPT レベル（手法 2）			
	J	S	J 準拠	S 準拠	平均	段階化	J 準拠	S 準拠	平均	段階化
CCH02	239	68	3	2.5	2.75	N3	2.22	2.04	2.13	N2
CCH03	248	77	3	2.6	2.80	N3	2.04	1.32	1.68	N2
CCH06	231	83	3	1	2.00	N2	2.38	0.84	1.61	N2
CCH07	287	81	2	1	1.50	N2	1.26	1.00	1.13	N1
CCH08	253	79	2	2.5	2.25	N2	1.94	1.16	1.55	N2
CCH09	294	85	2	1	1.50	N2	1.12	0.68	0.90	N1
CCH10	210	63	3	2.5	2.75	N3	2.80	2.44	2.62	N3
CCH11	261	80	2	2.5	2.25	N2	1.78	1.08	1.43	N1
CCH12	282	81	2	1	1.50	N2	1.36	1.00	1.18	N1
CCH13	227	77	3	2.5	2.75	N3	2.46	1.32	1.89	N2

注：表中、J、S、J 準拠、S 準拠、平均、段階化は、それぞれ、J-CAT 総合スコア、SPOT90 スコア、J-CAT 準拠の JLPT レベル推定値、SPOT 準拠の推定値、レベル推定値の平均、段階化されたレベル推定値を表す。

たとえば、CCH02 を例にすると、手法 1 では、平均レベル値が 2.75 で、段階は N3 相当となるが、各々のテストスコアを連続尺度である詳細スコア値に置き換えてから平準化する手法 2 では、平均レベル値が実際には 2.13 で、N2 相当であることがわかる。全体として、手法 1 ではランクを実際より低めに推定してしまうが、手法 2 ではそうした問題が解消されていることがわかる。

以上をふまえ、I-JAS for CIA では、上記の手法 2 によって合成した習熟度指標をレベル分類の主基準に採用し、主成分得点は参考資料として掲載する。なお、J-CAT スコアの換算に関して、回帰式によると、テストスコアが 350 点で JLPT レベル推定値が 0 になる。ゆえに、351 点を超える高いスコアを取った 3 名（KKR43、CCM51、KKD28）のレベル推定値はマイナスとなるが、連続尺度の性質をふまえ、今回はそのまま計算している。

2.5　新しい合成レベル値に基づく学習者の習熟度分類

以上をふまえ、1,000 名の学習者の各々について、2 種のスコアに基づき、JLPT レベル推定値を計算した。調査地別・母語別にまとめた結果は表 6 の通りである。

表 6　習熟度レベル別の人数とファイル数

	コード	N0	N1	N2	N3	N4	N5	N6	N	F
調査地	CCH		9	27	14				50	350
	CCM	1	12	26	10	1			50	350
	CCS	3	20	23	4				50	350
	CCT		16	28	6				50	350
	EAU			8	12	3			23	161
	EGB			3	10	4	2		19	133
	ENZ		1	2	10	7			20	140
	EUS		1	5	10	14	8		38	266
	FFR			4	30	15	1		50	350
	GAT			13	24	2			39	273
	GDE		1	2	6	2			11	77
	HHG		9	19	20	2			50	350
	IID			9	31	10			50	350
	JJC		8	17	17	6	1		49	343
	JJE			14	29	7	1		51	357
	JJJ	50							50	350
	JJN		4	18	9	10	7	2	50	350
	KKD	6	17	14	7	4			48	336
	KKR	6	24	19	3				52	364
	RRS		5	17	21	7			50	350
	SES			3	16	27	4		50	350
	TTH	1	6	20	20	3			50	350
	TTR		2	8	22	10	8		50	350
	VVN			11	28	11			50	350
母語[3]	CHN 中国語	4	57	104	34	1			200	1400
	DEU ドイツ語		1	15	30	4			50	350
	ENG 英語		2	18	42	28	10		100	700
	ESP スペイン語			3	16	27	4		50	350
	FRA フランス語			4	30	15	1		50	350
	HUN ﾊﾝｶﾞﾘｰ語		9	19	20	2			50	350
	IDN ｲﾝﾄﾞﾈｼｱ語			9	31	10			50	350
	KOR 韓国語	12	41	33	10	4			100	700
	RUS ロシア語		5	17	21	7			50	350
	THA タイ語	1	6	20	20	3			50	350
	TUR トルコ語		2	8	22	10	8		50	350
	VVN ベトナム語			11	28	11			50	350
	XJP (国内学習者)		12	49	55	23	9	2	150	1050
	JPN 日本語	50							50	350
N		67	135	310	359	145	32	2	1050	NA
F		469	945	2170	2513	1015	224	14	NA	7350

注：N は人数、F はファイル数（絵描写（D）ファイルは除く）を示す。

[3] 母語については、英語（ENG）を除き、調査地コードと混同しないよう、当該言語が話されている国の国際 3 文字コードで表記している。

以下、表 6 について、4 点補足する。1 点目は母語話者の位置づけである。本研究では N1 を超えるレベルを N0 と呼ぶため、母語話者も、少数の例外的に高い習熟度を持つ学習者とともに N0 に分類している[4]。

2 点目は、国内学習者の扱いについてである。国内学習者は、国内教室環境学習者（JCR）と国内自然環境学習者（JNR）に分けられるが、これらは、母語に基づく区分ではない。そこで表 6 では、2 群を併せて国内学習者（XJP）とする。なお、X は特定の母語を持たないことを示す。

3 点目は、I-JAS for CIA が対象とするデータ範囲についてである。I-JAS では、調査地により、ストーリーテリング 1（ST1）、ストーリーテリング 2（ST2）、対話（I）、ロールプレイ 1（RP1）、ロールプレイ 2（RP2）、絵描写（D）、ストーリーライティング 1（ST1）、ストーリーライティング 2（ST2）の 8 タスクを実施した群と、絵描写以外の 7 タスクを行った群が存在するが、群間の相互比較の目的に沿い、I-JAS for CIA は、絵描写以外のタスクを対象範囲とする。これにより、全群 7 タスクとなり、ファイル総数は 1,050 人×7=7350 本となる。なお、絵描写を実施したのは 757 名であるため、絵描写（D）を加えると、総ファイル数は 7,350＋757＝8,107 本となる。

4 点目は、CIA の対象となしうる群の特定である。CIA で群間比較を行うには、各群が、個体差を超えた何らかの傾向を示していることが前提となる。そこで、分析にかける 1 群の最低人数を仮に 5 人と決め、これに満たない箇所は網掛けで示す。これにより、習熟度を統制した CIA を行う場合に対象にできるデータ範囲がわかる。母語について言うと、12 の言語種すべてについて横断的に議論したい場合は N3 を対象にすべきである（水色部分）。それ以外だと、いくつかの母語で欠損が起こる。また、N0～N6 の 7 段階の習熟度の幅の中で、できるだけ広い範囲で発達過程を議論したいのであれば、韓国語話者（L0～L4 の 4 段階）、ロシア語話者（N1～N4 の 4 段階）、英語話者（N2～N5 の 4 段階）などを選べばよい。また、母語や学習状況（教室か自然か）が混在しているという制約はあるが、国内学習者を選べば、N1～N5 の 5 段階の比較が可能になる。

従来、十分に活用できていなかった 2 種のテストスコアを JLPT レベルという共通尺度を使って合成することで、表 6 に示すような習熟度を基準とした全学習者の分類が可能になり、これらを組み合わせることで、母語話者・学習者間比較、母語を異にする学習者間比較の両面において、研究者の関心をふまえたさまざまな CIA が実践できるようになるものと思われる。

３．I-JAS for CIA の構築

以上で、新しい習熟度基準に基づき、学習者の群化ができたわけだが、この枠組みで CIA を実施するには、該当する学習者の産出テキストだけをまとめて処理することが必要になる。

だが、現行のダウンロード版テキストは、前述のように、(1)習熟度や母語による分類がなされていない、(2)フォルダ構造が複雑でデータが取り出しにくい、(3)テキストファイルの中に行コードと産出テキストが混在しており、ファイルのまま解析が行いにくい、といったいくつかの制約があり、習熟度・母語別に適切なデータだけを取り出すことはきわめて困難になっている。

そこで筆者は、表 7 の手順で作業を進め、絵描写（D）を除く全産出データと話者属性データのすべてを 1 枚のワークシートに集約し、習熟度を統制した CIA の実施を可能にする I-JAS for CIA として整備することとした。

---

[4] こうした扱いは、「母語話者」を学習者とは異なる特別な存在と見なさないという理念に基づく。理論的には、N0 から外れる「母語話者」が存在する可能性もありうるが、その点を検証するデータはないため、今回は 50 名の母語話者を一律で N0 としている。

表 7　I-JAS for CIA の整理手順

手順	差業種別	主な作業内容
1	ファイルの集約	ダウンロード版テキストの全体を 1 か所に集約
2	ファイルのマージとエクセルへの転記	全テキストファイルを単一ファイルに統合し、単一のエクセルシートに転記
3	行コードからの属性情報の取り出し	行コードから、調査地・ID・タスク・話者の情報を独立したセルに取り出す
4	母語属性情報の追加	母語を新たな属性コードとして情報付与
5	習熟度情報の追加	習熟度を新たな属性コードとして情報付与
6	テキストクリーニング	開発者によって付与された編集情報を削除した列を別途新設
7	インタビュイー産出取り出し	話者コードを用い、インタビュイー産出行のみを新しいシートに複写

以下、各々の手順について実際の作業を示す。

3.1　ファイルの集約

I-JAS のテキストデータを入手するには、「中納言」に示されたリンクから「データ配布 1」サイトを開き、対面調査（プレインテキスト）フォルダを選び、そこに表示される調査地別の 24 種の圧縮ファイルを 1 つずつダウンロードする必要がある（本稿執筆時点で、一括ダウンロードはできない）。

図 4　I-JAS の「データ配布 1」サイトの構造

ダウンロードしたデータを解凍すると、24 種のフォルダができるが、その各々に、8 つのサブフォルダがあり、実際のテキストはその中に収納されている。

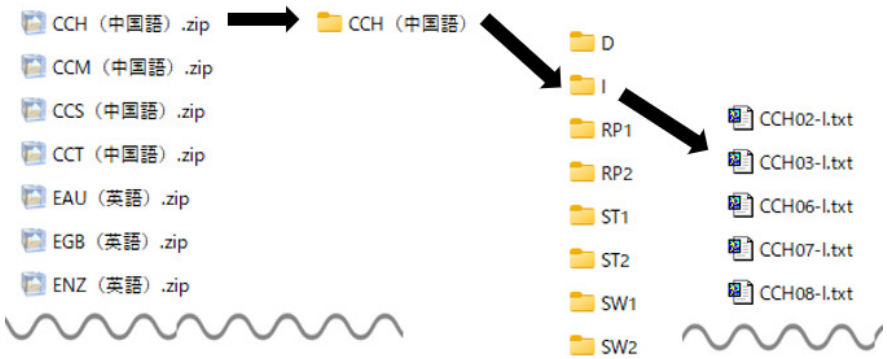

図 5　ダウンロードデータからテキストファイルへのアクセス

上記は、ダウンロードした圧縮ファイルの中から CCH を解凍し、CCH フォルダ内の全サブフォルダを表示させた後、その中の対話（I）サブフォルダを開いて、中にあるテキストファイルを確認するステップを示している。このような形ですべてのテキストファイルを取り出すには、160 以上のサブフォルダを 1 つ 1 つ開けていかなければならない。こうした作業は煩瑣であるのみならず、必要なファイルを見逃すリスクもある。そこで、以下では、全ファイルをまずは一括的に集約することとしたい。

作業手順として、デスクトップ上に新しいフォルダを作り（たとえば、I-JAS 24 folders と命名）、その中に、解凍済みの 24 フォルダをすべて入れる。その上で、新しく作ったフォルダを開き、上部にある検索ボックスに「.txt」と入力する。これで、フォルダ内にある全テキストファイル（D を含む全 8,107 本）が検索結果画面に一覧表示される。なお、環境にもよるが、この作業に数分以上を要する場合もある。画面下部の左端に表示される「n 個の項目」の数が 8,107 になれば検索は終了している。

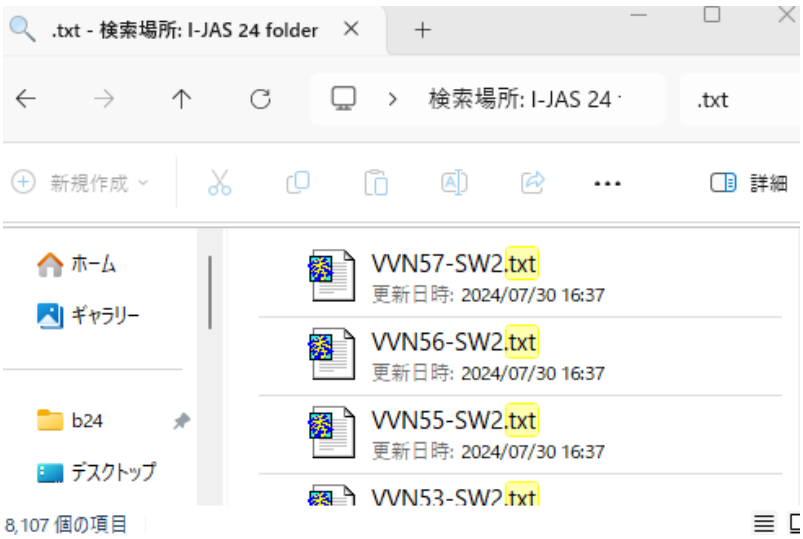

図 6 単一フォルダ内にあるテキストファイル全部を検索で表示させ、移動する

その後、新しいフォルダ（たとえば all texts と命名）を作り、図 8 で表示されている全 8,107 ファイルを新フォルダにコピーする。ファイル数が多いため、この作業にも、数分以上かかる場合がある。

次いで、絵描写タスクのデータを除くため、all texts フォルダで D.txt を検索し、検索結果に表示された 757 ファイルを別フォルダ（たとえば D all texts と命名）に移動する。これで、all texts フォルダには、I-JAS for CIA が対象とする 8,107—757=7,350 本のみが残ることになる。

3.2 ファイルのマージとエクセルへの転記

続いて、7,350 本のテキストファイルを単一のテキストファイルにマージする。マージは OS のコマンドでも実施可能だが、今回は、Text Coupler というツールを用いる。このツールは、マージの前にファイルの並び順の確認や、ファイル間のセパレータの有無、ファイルのエンコード形式の確認・強制指定などができる。

まず、7,350 本のファイルをツールに読み込ませる（筆者の環境ではこのプロセスが完了するまで 10 分以上の時間を要した）。

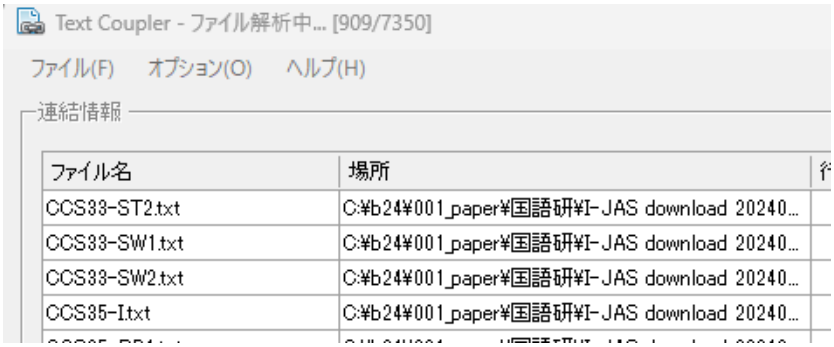

図 7　Text Coupler へのファイルの読み込み（読み込み中の画面）

その後、ファイル名で全体をソートし、ファイル間の切れ目は 0 行として（連結するため）、ファイル形式を UTF-8 と指定して全体を単一のテキストファイルとして出力する。

7,350 本をマージしたテキストファイルは、サイズが 50MB を超える大きなものになるが、これを開き、全体をエクセルのシート上にコピーする。テキストファイルには行コードとテキストが混在しているが、CSV 形式で保存されているため、エクセルにコピーすると、行コードと産出部が自動的に切り分けられ、別の列として記載される。

	A	B
1	CCH02-I-00010-C	ＩＤお願いします
2	CCH02-I-00020-K	ＣＣＨ０４です
3	CCH02-I-00030-C	はーい、えっ？
4	CCH02-I-00040-K	あぜ
5	CCH02-I-00050-C	もう一回言っ（ゆっ）って
6	CCH02-I-00060-K	えＣＣＨ０２です
7	CCH02-I-00070-C	はい０２ですね、あの今日は本
8	CCH02-I-00080-K	あー四年生になりなったら、ん
9	CCH02-I-00090-C	あそうですか、じゃあちょっと
10	CCH02-I-00100-K	はい

図 8　マージファイルをエクセルに転記（行コードとテキストは別の列に分かれる）

今回の処理では、エクセル上で、全体が 557,315 行となった。なお、筆者が使用したエディタ（「秀丸」）上の行数は、折り返しがあるため、77 万行程度となっている。エディタによっては、事前に、行数の許容値を引き上げておかないと全体を 1 ファイルで管理できない場合があるので注意されたい。

３.３　行コードからの属性情報の取り出し

続いて、対象データの検索や抽出を行いやすくするため、行コードに埋め込まれている基本的な属性情報を取り出す。I-JAS の行コードの構造は下記のようになっている。

CCH02　-　ST1　–　00020　-　K
|ID|　|タスク|　|行番号|　|発話者|

図 9　行コードの構成概念図

ここからまず、調査地（ID の左側 3 文字）、参加者の ID、タスク、発話者（インタビュ

ワーが C、インタビュイーが K）コードを取り出す。調査地と ID の取り出しにはエクセルの LEFT 関数を（左側 3 または 5 文字を取り出す)、発話者の取り出しは RIGHT 関数を（右側 1 文字を取り出す)、タスクの取り出しには MID 関数（冒頭 7 文字目から 3 文字を取り出す）を使用する。なお、タスクは上記の図 9 ように原則 3 文字であるが、対話（I）のみ I の 1 文字となっているため、すべて INT（interaction の頭文字）に置換してから処理を行う。

調査地	ID	発話者	タスク	コード	テキスト1
CCH	CCH02	C	INT	CCH02-IN	ＩＤお願い
CCH	CCH02	K	INT	CCH02-IN	ＣＣＨ０４
CCH	CCH02	C	INT	CCH02-IN	はーい、え
CCH	CCH02	K	INT	CCH02-IN	あぜ
CCH	CCH02	C	INT	CCH02-IN	もう一回言
CCH	CCH02	K	INT	CCH02-IN	えＣＣＨ０
CCH	CCH02	C	INT	CCH02-IN	はい０２で

図 10　調査地・ID・発話者・タスク情報の抽出

３.４　言語コードの付加

以上で、I-JAS for CIA データシートの基本枠が完成したわけだが、「多言語母語の学習者横断コーパス」という名称が示すように、このコーパスは、学習者の「母語」に基づく比較を前提として作成されたものと思われる。

にもかかわらず、おそらくは調査地単位で段階的にデータ公開をしてきたことの名残によるものか、I-JAS のデータは、オンライン版・ダウンロード版とも、基本的には、母語でなく、調査地の単位で整理されているようである。

「中納言」には、「同じ言語環境（母語）の調査地 ID の違いは地域・教育機関の違いであり、調査条件が異なるものでは」ない旨の注記が小さく書き込まれているが、検索システム上では、一部の例外を除くと、CCH、CCM といった調査地モジュールごとに検索対象を選択する仕様になっており、「中国語」や「韓国語」といった母語単位で選択することはできない。また、ダウンロード版は、調査値ごとに独立したフォルダとなっている。これらの点で、コーパスの本来の狙いと、現在の検索システムやダウンロード版のデータ構造の間には、やや乖離があるかもしれない。

実際、調査地を言語研究の分類単位にすることには疑問が多い。「中納言」上でも、絵描写(D)タスクの有無という点から CCH と CCM は分かれているが、一方で、KKD と KKR、EGB と EUS、EAU と ENZ はまとめられており、そもそも、調査地別という単位で一貫しているわけではない。また、KKD や KKR などについて、調査地ごとの詳細は秘匿されており、これらを区別して分析するメリットはほとんど見当たらない。

言語環境・調査地およびタスク

言語環境	海外																日本国内		
	中国語			韓国語	タイ語	ベトナム語	インドネシア語	英語		ドイツ語	フランス語	スペイン語	ロシア語	ハンガリー語	トルコ語	国内教室環境	国内自然習得	日本語母語話者
	中国D有	中国D無	台湾					D有	D無									
人数	50人	50人	100人	100人	50人	50人	50人	57人	43人	50人	50人	50人	50人	50人	50人	100人	50人	50人
調査地ID	CCH	CCM	CCS CCT	KKD KKR	TTH	VVN	IID	EGB EUS	EAU ENZ	GDE GAT	FFR	SES	RRS	HHG	TTR	JJC JJE	JJN	JJJ
ST1	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

図 11　I-JAS のデータ構造（「中納言」検索インタフェース）

そこで、I-JAS for CIA では、I-JAS の本来の狙いをより忠実に反映した検索を推奨する観点から、現在の調査地コードに加え、新たに母語 1（日本語で表記）と母語 2（略号で表記）という列を設けることとした。2.5 節の注釈でも述べたように、後者には、英語（ENG）を除き、当該言語が主に話されている国の国際 3 文字コードを使用する。国内教室環境学習者（JJC/JJE）と国内自然習得学習者（JJN）は、母語で区別されていないため、XJP という特殊コードを付与して 1 つにまとめる。

母語1	母語2	調査地	ID	発話者	タスク	コード
中国語	CHN	CCH	CCH02	C	INT	CCH02-IN
中国語	CHN	CCH	CCH02	K	INT	CCH02-IN
中国語	CHN	CCH	CCH02	C	INT	CCH02-IN
中国語	CHN	CCH	CCH02	K	INT	CCH02-IN

図 12　母語コード（母語 1、母語 2）の追加

これにより、母語 1 ないし母語 2 の列にフィルタをかけることで、中国語話者の産出全体、英語話者の産出全体、といった抽出が簡単にできるようになる。

３.５　習熟度情報の付加

すでに述べたように、CIA の妥当性を高めるには、習熟度の統制が不可欠である。そこで、2.4 節で示した処理によって得られた習熟度情報を I-JAS for CIA のシートにも追加しておくこととしたい。

作業手順としては、同じシート上に、別途作成した学習者 1,000 名分の習熟度情報のリストを貼り付け、VLOOKUP 関数で、リストから J-CAT 総合スコア、SPOT90 スコア、J-CAT 準拠レベル値、SPOT90 準拠レベル値、平均レベル値、段階化レベル値の情報を抽出する。

このうち、「段階化レベル値」とは、前述のように、2 種のテストスコアを、回帰式によって連続尺度化された JLPT レベル推定値に置き換えて平均を取り、小数点第 1 位で四捨五入を行って N0（平均レベル値が 0.00～0.49）～N6（5.50～6.49）の 7 段階に整理した値である。段階化レベル値は、CIA で習熟度を統制する際の基本的な指標として使えるもの

である。なお、JLPT レベルと対応していることがわかるよう、数値だけでなく N を付して表示する。

また、参考資料として、J-CAT の 4 種のセクションスコアと SPOT90 のスコアを合成した主成分得点（2.3 節）の情報も併せて表示する。

Jスコア	Sスコア	PCA	J準拠レ	S準拠レ	平均レイ	段階化レ	ID
239	68	1.05	2.22	2.04	2.13	N2	CCH02
248	77	1.59	2.04	1.32	1.68	N2	CCH03
231	83	1.47	2.38	0.84	1.61	N2	CCH06
287	81	2.86	1.26	1	1.13	N1	CCH07
253	79	1.84	1.94	1.16	1.55	N2	CCH08

図 13　7 種の習熟度情報の追加

筆者としては、段階化レベルの使用を推奨するが、ユーザーは、これ以外であっても、自身の研究目的により、任意の習熟度情報を基準として、学習者の群化や習熟度の統制を行える。

3.6　テキストクリーニング

以上で、各種の属性情報が I-JAS for CIA データシートに取り込めた。最後に、産出テキストに注目しよう。ここで留意しておくべきは、I-JAS のダウンロード版テキストには、開発者による修正や補注など、実際にインタビュイーが産出したテキスト以外の内容が各種のカッコ書きで加えられているということである。下記は CCH03 の対話（I）データ中の例である。

(1) えのー、えっとー、えー【大学名 1】、大学は、あー二つの、アパートがあります
(2) たぶん、十、じゅうふん（十分）ぐらい・・・
(3) えー、自転車、いちと（一度）、…　もう、乗ーれてませ（乗れてません）
(4) あく、あーあるい、ええ、あーあー、歩くと〈うん〉、えー、えー
(5) この（連体詞）、かん、あーキャンパスから…
(6) え、小さいの（過剰使用）頃…
(7) えー、あー、彼はあの、か、あー、あー、彼の授業、に…、えー｛笑｝わ、私ぬ、私があの自分、あー自分で…

(1)は固有名詞をマスキングするためのもので、大学名のほか、【人名】【学部名】【施設名】などのバリエーションがある。(2)と(3)は学習者の言い間違えの修正例である。(4)はインタビュワーによる相槌や合いの手などの記載である。(5)と(6)は品詞や言語学的な情報の追加である。(7)は非言語的発声の追加であり、このほか、｛咳｝｛ため息｝などのバリエーションがある。

これらはデータの解釈に有益な情報を与えてくれるものだが、学習者が実際に産出したテキストに限定して語数カウントを行うといった場合であれば、こうした加工がないほう

が望ましい場合も想定される。そこで、産出列の全体をテキストエディタにコピーし、正規表現を用い、¥（.*?¥）→空白、という一括置換を行うことで、208,784 個の（　）とその内部を消去した。続いて、同様の手法で、188,509 個の〈　〉、40,875 個の｛　｝、3,908 個の【　】とその内部を削除した。なお、類似した正規表現として、他に　¥（.*¥）ががあるが、これだと、1 行中に複数の括弧が入っている場合、最初の開き括弧の後がすべて削除されてしまうので注意が必要である。

こうして、挿入を除去したテキストを作成し、テキスト 1（クリーニング前のテキスト）とともに、テキスト 2（クリーニング後のテキスト）をデータシートに加えることとした。

テキスト1	テキスト2
ＩＤお願いします	ＩＤお願いします
ＣＣＨ０４です	ＣＣＨ０４です
はーい、えっ？	はーい、えっ？
あぜ	あぜ
もう一回言っ（ゆっ）って	もう一回言って
えＣＣＨ０２です	えＣＣＨ０２です
はい０２ですね、あの今日は	はい０２ですね、あの今日
あー四年生になりなったら、	あー四年生になりなったら
あそうですか、じゃあちょっ	あそうですか、じゃあちょ
はい	はい
あの今日はうちからここまで	あの今日はうちからここま
つ、あるって（歩いて）、〈	つ、あるって、歩く、え

図 14　クリーニング後のテキスト（テキスト 2）の追加[5]

なお､学習者が産出した厳密な語数を議論する場合はクリーニングを行ったテキスト 2 の使用が好ましいが、その他の分析目的に関しては、両者を比較して適切なほうを選ぶべきである。たとえば、上記引用のテキスト 1 の最終行の「あるって」は「歩いて」の発音がやや曖昧だった例だが、テキスト 2 で解析すると、「有るって」と解されてしまい、学習者が言おうとしていた「歩いて」は出てこなくなってしまう。この意味で、語の内容分析などを行う場合は、テキスト 1 を使うほうが合理的だろう。

３.７　付加シートの作成

I-JAS for CIA をより使いやすくするため、以上で作成したシート（base と命名する）以外に、3 枚のシートを追加する。

1 枚目はこのデータシートを初めて使うユーザーを対象として、データシート開発の概要をまとめたシート（about）、2 枚目は base シートにあるデータのうち、インタビュイー側

[5] 図 14 のテキスト 1 の 5 行目は、「言って」に発音情報を追加したものだが、ここは「言（ゆ）って」もしくは、「言って（ゆって）」とあるべきだったかもしれない。促音の「っ」が 1 つ余剰のように見える。

産出（291,161 行）のみを取り出したシート（learner turn[6]）である。base シートは、全体の行数が多く、検索などにも時間がかりがちなので、場合によっては、learner turn シートを使うほうが便利だろう。また、3 枚目は学習者の習熟度情報を記録したシート（proficiency）である。

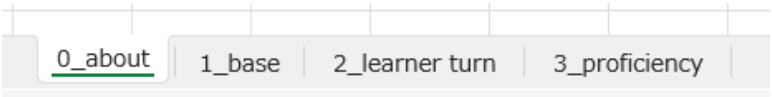

図 15 4 種のシートの切り替えタブ

全体で 4 種のシートには、インデックスを付しているので、ユーザーは適宜切り替えて使用することができる。

４．I-JAS for CIA の活用

４.１　I-JAS for CIA の使い方

以上で完成した I-JAS for CIA の base シート（learner turn も同じ）は、下記のような列の構成になっている。

習熟度							属性							産出	
Jスコア	Sスコア	PCA	J準拠レイ	S準拠レイ	平均レイ	段階化レ	母語1	母語2	調査地	ID	発話者	タスク	コード	テキスト1	テキスト2
239	68	1.05	2.22	2.04	2.13	N2	中国語	CHN	CCH	CCH02	K	INT	CCH02-INT-00020-K	ＣＣＨ０４です	ＣＣＨ０４です
248	77	1.59	2.04	1.32	1.68	N2	中国語	CHN	CCH	CCH03	K	INT	CCH03-INT-00020-K	こんにちは	こんにちは
231	83	1.47	2.38	0.84	1.61	N2	中国語	CHN	CCH	CCH06	K	INT	CCH06-INT-00020-K	ＣＣＨ６番です	ＣＣＨ６番です
287	81	2.86	1.26	1	1.13	N1	中国語	CHN	CCH	CCH07	K	INT	CCH07-INT-00020-K	はいＣＣＨ、あー０７です	はいＣＣＨ、あー０７です
253	79	1.84	1.94	1.16	1.55	N2	中国語	CHN	CCH	CCH08	K	INT	CCH08-INT-00020-K	はい、ＣＣＨ０８です	はい、ＣＣＨ０８です

図 16　base シートの列内容（全体）

全体像がわかりやすいよう、習熟度・属性・産出という 3 つの区分を設け、それぞれに下位区分を設定し、その各々に検索用のフィルタを設定している。

習熟度						
Jスコア	Sスコア	PCA	J準拠レイ	S準拠レイ	平均レイ	段階化レ
239	68	1.05	2.22	2.04	2.13	N2
248	77	1.59	2.04	1.32	1.68	N2
231	83	1.47	2.38	0.84	1.61	N2
287	81	2.86	1.26	1	1.13	N1
253	79	1.84	1.94	1.16	1.55	N2

図 17　習熟度関連の列

[6] 本来は母語話者も含むので interviewee turn と呼ぶべきだが、ユーザーのわかりやすさのために learner tun としている。

属性						
母語1	母語2	調査地	ID	発話者	タスク	コード
中国語	CHN	CCH	CCH02	K	INT	CCH02-INT-00020-K
中国語	CHN	CCH	CCH03	K	INT	CCH03-INT-00020-K
中国語	CHN	CCH	CCH06	K	INT	CCH06-INT-00020-K
中国語	CHN	CCH	CCH07	K	INT	CCH07-INT-00020-K
中国語	CHN	CCH	CCH08	K	INT	CCH08-INT-00020-K
中国語	CHN	CCH	CCH09	K	INT	CCH09-INT-00020-K

図 18　属性関連の列

産出	
テキスト1	テキスト2
ＣＣＨ０４です	ＣＣＨ０４です
こんにちは	こんにちは
ＣＣＨ６番です	ＣＣＨ６番です
はいＣＣＨ、あー０７です	はいＣＣＨ、あー０７です
はい、ＣＣＨ０８です	はい、ＣＣＨ０８です

図 19　産出テキスト関連の列

ユーザーは、習熟度や属性のフィルタを組み合わせることで、たとえば、N3 と N4 レベルの、韓国語母語の、学習者（発話コード K）の、ロールプレイ 1 の産出や、N1 レベルの、英語母語の、学習者の、ストーリーライティング 2 の産出などを、簡単に取り出すことができる。また、調査地のフィルタを使えば、中国語話者のうち、大陸学習者（CCH，CCM）と台湾の学習者（CCS と CCT）を区別することもできる。下記は「段階化レベル」のフィルタで、N3 と N4 を指定した際の画面である。

図 20　レベルフィルタ（N3 および N4 レベルのみを取り出す例）

こうして、習熟度と属性のフィルタを組み合わせ、任意の話者・タスクの産出だけを表示させた後は、2 種の研究の進め方がある。1 つ目は、表示されているテキスト部（テキスト 1 またはテキスト 2）を列ごとコピーして新しいテキストファイルを作り、それを「web 茶まめ」などで形態素解析し、形態素解析済みのファイルを Antconc などのコンコーダンサで分析する方法である。このやり方だと、精密な語彙頻度分析や、コンコーダンサの機能を用いた共起語分析、連語抽出、特徴語抽出等を行うことができる。

もう 1 つは、フィルタを組み合わせて特定の話者・タスクの産出のみを表示させた段階で、テキスト 1 またはテキスト 2 にあるフィルタボタンを押して「テキストフィルタ」を開き、「指定の値を含む（A）」から、関心対象の語や表現を入力し、当該形を含むすべての行を自動抽出して観察する方法である。

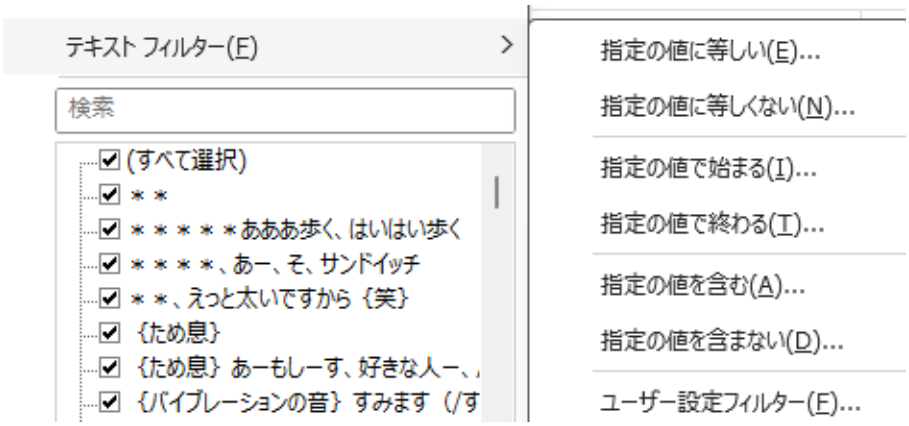

図 21　産出列のフィルタ

たとえば、「していて」を含む、というように抽出条件を指定すると、「していて」を含む行だけが抽出される。

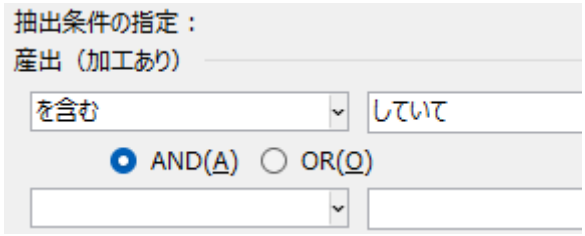

図 22　「していて」を含む行を検索

また、2 条件を同時に指定することもできる。下記のように指定すれば、「します」または「しました」を含む行が一気に抽出できる。

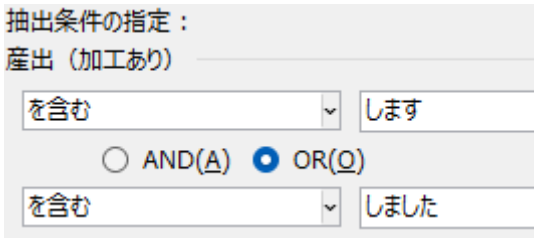

図 23　「します」または「していた」を含む行を検索

こうすることで、たとえば、N3 と N4 の学習者の「します・しました」の用例数を比較したり、インドネシア母語話者と日本語母語話者（ともに 50 人）の「します・しました」の用例数を比較したりすることも可能になる。ただし、厳密に言えば、テキストフィルタで取り出せるのは行数であり、1 行中に複数の「します」や「しました」が出ている場合は、行数と頻度が一致しないことになる点には注意が必要である。

４.２　I-JAS for CIA の活用例：習熟度を統制した学習者母語の相互関係調査

前述のように、条件に合致する用例を抽出するだけであれば、I-JAS for CIA のシートだけでもかなりのことができる。しかし、厳密な語数を調査したい場合や、テキスト全体に対して高度な解析を行いたい場合は、I-JAS for CIA からの抽出結果を別途テキストファイル

に貼り付けて保存し、独立した分析を行うことになる。

本節では、習熟度を統制した計量的なデータ比較の一例として、12 種の母語背景を持つ N3 レベル[7]の学習者と母語話者を対象に、絵描写を除く全タスクの産出総体（1,284,950 語）における動詞頻度を手掛かりとした話者群分類の結果を示したい。研究上の関心は、12 種の話者群が、母語の言語的な親疎関係に沿って分類されるかどうかを確認することにある。

手順として、まず、I-JAS for CIA の learner turn シート上で、「段階化レベル」を N3 に固定した上で、「母語 1」列のフィルタにより、12 言語を順に指定し、その都度、抽出されたテキスト 1 の列をコピーしてテキストファイルとして保存する。その後、「段階化レベル」を N0 にして、「母語 1」列のフィルタにより「日本語」を設定して、同様に抽出されたテキスト 1 列全体を保存する。これで、13 種のテキストファイルが作られる。ファイル名には母語コードを付けておく。

さて、日本語テキストデータから語彙頻度を取り出すには、形態素解析が必要となる。この目的には、たとえば「web 茶まめ」などを利用することができるが、今回は、筆者の研究室で開発・リリースした English/Japanese Word Frequency Table Generator（EJWFTG）を利用する。EJWFTG は Google Colab 上のオンラインツールで、複数ファイルをアップロードすると、自動で、UniDic/MeCab による解析が行われ、各語についてファイルごとの頻度を一覧表示する統合頻度表が出力される。なお、EJWFTG の頻度表は、表層形、表層形＋品詞、語彙素、語彙素＋品詞の 4 モードで出力される（詳細は、石川，2024；石川，forthcoming を参照されたい）。

今回は、語彙素＋品詞のデータを用い、文産出の根幹を担う動詞上位 100 語の頻度を抽出する。

word	CHN	DEU	ENG	ESP	FRA	HUN	IDN	JPN	KOR	RUS	THA	TUR	VNM
為る-動詞	1554	1183	1662	645	1092	824	1134	3471	416	836	784	763	1285
有る-動詞	762	701	895	348	687	518	793	1958	172	568	521	619	601
居る-動詞	772	791	811	343	790	466	753	1585	185	448	491	640	634
言う-動詞	253	427	481	172	250	204	314	3810	91	203	238	305	211
思う-動詞	392	491	920	204	291	372	641	1506	119	453	338	521	397
行く-動詞	390	351	652	169	347	287	321	1310	105	301	230	358	377
見る-動詞	325	321	500	196	282	230	412	773	90	280	208	308	264
成る-動詞	245	141	394	137	190	189	366	986	76	216	152	251	281
来る-動詞	250	269	242	112	230	170	227	856	86	217	122	297	178
食べる-動詞	274	212	303	99	223	89	231	414	67	126	163	158	188
出来る-動詞	163	160	255	137	198	120	206	509	64	171	83	176	170
入る-動詞	283	157	239	90	127	84	226	423	50	87	104	178	185
分かる-動詞	174	200	286	100	210	117	151	361	24	152	95	145	145
作る-動詞	136	184	167	72	128	82	150	247	16	146	100	190	192

図 24　EJWFTG の語彙素＋品詞データより動詞のみを抽出（一部）

こうして、話者 13 群を変数、高頻度動詞上位 100 語をケースとする頻度表を作成した。該当する動詞は下記である。

表 8　分析に使用した上位 100 動詞

動詞（語彙素）頻度上位 100 種（頻度降順で記載）
為る、有る、居る、言う、思う、行く、見る、成る、来る、食べる、出来る、入る、分かる、作る、済む、住む、話す、仕舞う、働く、遣る、知る、寝る、持つ、起きる、聞く、出る、返る、御座る、入れる、呼ぶ、開ける、取る、違う、使う、読む、呉れる、貰う、買う、教える、遊ぶ、考える、つく、知れる、忘れる、覚える、願う、掛かる、会う、上げる、探す、書く、置く、終わる、上る、乗る、頑張る、歩く、始める、気付く、頂く、下さる、変わる、付く、

[7] 本論文 2.5 節で述べたように、12 種の母語背景を持つ学習者すべてを横断的に比較する場合は、「段階別レベル」を N3 にする必要がある。

生まれる、空く、見付ける、切る、聞こえる、選ぶ、死ぬ、変える、見える、飲む、過ぎる、出す、困る、掛ける、怒る、飛ぶ、待つ、着く、開く、為さる、受ける、止める、手伝う、飛び出す、因る、戻る、押す、走る、残る、習う、出掛ける、決める、続ける、喋る、歌う、休む、慣れる

上記の頻度表に対して変数クラスター分析（距離計算は 2-2r、融合後の距離は Ward 法による）を実施したところ、以下の結果を得た。なお、クラスター分析では、図の左から右側に向かい、関係の近いものから順に融合が進んでいき、全体が 1 つの大クラスターに統合された段階で分類は終了する。この過程が樹形図として出力される。

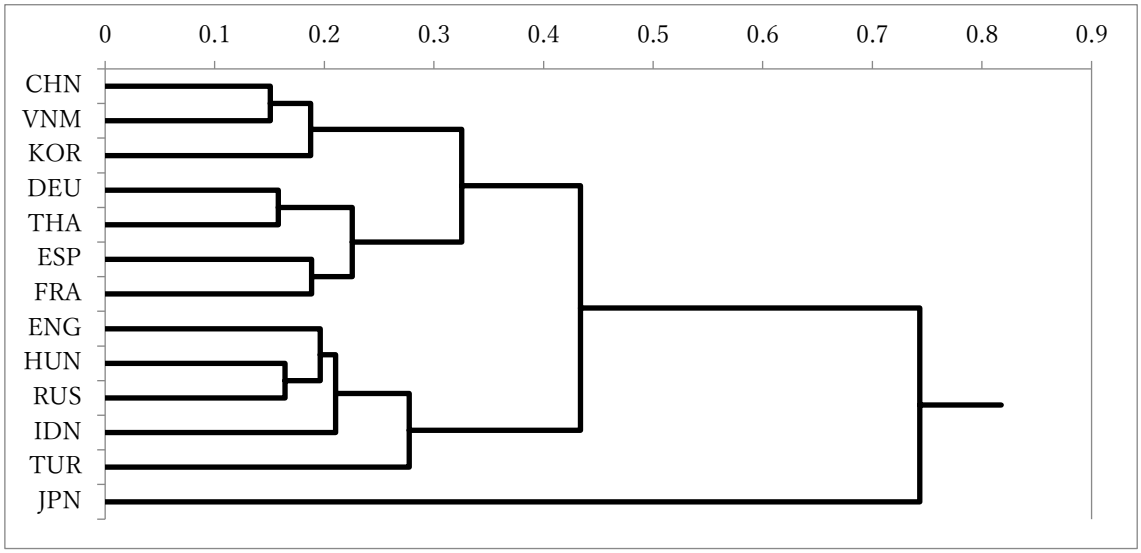

図 25 変数クラスター分析による樹形図（動詞に基づく話者群の分類）

図 25 より、発話および作文中での動詞使用に関して、(1)学習者と母語話者は完全に分離されること（両者の融合が起こるのは最も最後の段階である）、(2)中国語話者とベトナム語話者、ドイツ語話者とタイ語話者、スペイン語話者とフランス語話者、ハンガリー語話者とロシア語話者が、それぞれ相対的に近い関係にあること、(3)学習者は英語・ハンガリー語・ロシア語・インドネシア語・タイ語話者群と、それ以外の群に大別されること、などが明らかになった。

なお、迫田他（2020）には、これら 12 言語の言語類型が記載されている（p. 16）。12 言語の類型はさまざまだが、今回、近い距離にあると判定されたペアのうち、スペイン語とフランス語は印欧語族（イタリック語派）で語族も近い。一方、今回のデータで近い距離にあるとされたその他の組み合わせについては言語類型上の親近性が確認されなかった。少なくとも発話・作文における L2 日本語の動詞使用について言えば、いわゆる L1 の言語類型の影響は明白とは言いにくいように思える。

言語類型論の立場から詳細な解釈を行うことは本稿の趣旨を超えるが、いずれにせよ、こうした議論が信頼性を持ちうるのは、事前に習熟度を揃えることで、習熟度の影響を統制しているからである。どのような目的の研究であれ、母語の比較を行う際には、「母語以外の条件を同一にする必要」があり、I-JAS の「詳細な習熟度を加味することで、信頼性の高い比較研究」が可能になるとされているが（迫田，2020, p.90）、I-JAS for CIA を使えば、こうした分析が平易に行える。この意味で、I-JAS for CIA は、I-JAS を用いた幅広い CIA 研究を触発するユニークな言語資源になりうる。

５．まとめ

以上、本稿では、I-JAS for CIA という新しいデータシートの作成過程と、その利用法について解説した。

まず、1 章では、I-JAS のダウンロードデータの制約について触れ、2 章では 2 種の習熟度スコアデータの合成方法を論じ、3 章では、新たに導入された習熟度指標に基づいてテキストを分類し、I-JAS for CIA として整理していく過程を示した。また、4 章では、I-JAS for CIA の活用法を解説した後、習熟度統制に基づく CIA 実践の一例として、12 の母語背景を持つ学習者群および日本語母語話者群を、高頻度動詞上位 100 語を手掛かりに分類する研究例を紹介した。

冒頭でも述べたように、I-JAS は、高度な CIA に対応する先進的な設計がなされているにも関わらず、習熟度データの複雑性や、ダウンロード版のテキストデータの扱いにくさから、本格的な CIA 研究の普及がなかなか進まないという実情にあった。I-JAS for CIA が、海外での学習者コーパス研究と直接的に連結しうる、新しいタイプの I-JAS 研究の普及の一助になればと願う。

I-JAS for CIA の一般公開については、今後、開発者チームと議論を進めていく予定である。また、将来の計画としては、I-JAS のフェースシートに含まれるその他の調査項目についても再構築を行い、データの構造や階層に大幅な整理を加えた上で、一般にも使いやすい形で、I-JAS for CIA のデータに追加していくことを構想中である。

謝　辞

本研究は、アジア圏における学習者コーパス構築の基盤整備を行う科研費プロジェクト 23H00641、および、日本語学習者コーパスのグローバルネットワーク構築を目指す科研費プロジェクト 19KK0055 の支援を得て実施したものである。また、本研究の一部は、国立国語研究所の共同研究プロジェクト「多世代会話コーパスに基づく話し言葉の総合的研究」（プロジェクトリーダー：小磯花絵）における成果の一部である。関係の各位、とくに、本邦における第 2 言語習得研究の裾野を大きく広げる I-JAS という素晴らしい言語資源を開発された迫田久美子氏の多大なるご尽力とご貢献に改めて敬意を表したい。

文　献

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*, 3980. https://doi.org/10.3389/fpsyg.2021.715843

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies, Lund, 4-5 March 1994* (pp. 37−51). Chartwell-Bratt.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7−24.

石川慎一郎（2024）「『森を見ながら木を見る』コーパス研究の意義 : 複数テキストから統合語彙頻度表を作成する EJWFTG の開発」『統計数理研究所共同研究リポート』*469*, 95−122. https://doi.org/10.24546/0100487709

石川慎一郎（forthcoming）「English/Japanese Word Frequency Table Generator（EJWFTG）を用いた日本語統合語彙頻度表の作成と活用」『計量言語学』*34*(6), 421−431.
迫田久美子・石川慎一郎・李在鎬（編著）（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・小西円・佐々木藍子・須賀和香子・細井陽子 (2016).「多言語母語の日本語学習者横断コーパス」『国語研プロジェクトレビュー』*6*(3), 93−110.

関連 URL

English/Japanese Word Frequency Table Generator（EJWFTG） bit.ly/EJWFTG