

統合型日本語語彙データベースJ-CLID:C4の構築

Development of the Multipurpose Japanese Vocabulary Database J-CLID:C4

呂 建 輝

LU Jianhui

岡山大学大学院社会文化科学研究科紀要
第58号 2024年12月 抜刷
Journal of Humanities and Social Sciences
Okayama University Vol.58 2024

統合型日本語語彙データベース J-CLID:C4の構築

呂 建 輝*

1. 日本語語彙データベースの現状と課題

日本語語彙のデータベースの開発が進んでおり、これらの電子リソースを活用した応用研究・開発研究も盛んになっている。しかし、データベースによって、語彙の判定基準も語彙情報（表記、読み、アクセントなど）も異なるため、どのデータベースの情報を信用すればよいか迷うことも少なくないだろう。データベースは人によって作られたものである以上、作成基準が異なるのも当然のことだが、よりデータの信用性を高めるためには、複数の視点から総合的に評価する必要がある。同じ言葉を、複数の辞書で調べてみれば、よりよい知見が得られるというようなものである。しかし、日本語語彙のデータベース間における対応関係に関する研究が少なく、データベースの信用性に対する検証は未だに不十分である。

また、現行の日本語語彙データベースは基本的に目的別に作られている。例えば、以下本稿で取り扱う「日本語教育語彙表」は日本語非母語話者の日本語語彙学習のために作られたもので、日本語非母語話者にとっての語彙難易度情報が入っている。「教科書コーパス語彙表」は日本語母語話者が小・中・高校で学ぶ日本語語彙をリスト化したもので、日本語母語話者にとっての語彙難易度情報や使用分野といった情報が入っている。もしこの2表間の横断検索ができれば、日本語母語話者と非母語話者の間の語彙難易度のズレを明らかにでき「やさしい日本語」の研究に貢献できる。さらに、専門分野情報を活かせば日本語非母語話者の専門教育にも役立つと考えられる。

このように、データベースの信用性を向上させるためにも、データベース間の横断検索により新たな知見を得るためにも、データベース間の連携が必要であると考えられる。本研究では、日本語非母語話者対象の日本語教育や日本語母語話者対象の国語教育でよく使用される日本語語彙データベースを一表に統合させ、統合型の日本語語彙データベースを構築する。

2. データベース統合の手順

2. 1. 統合の対象となるデータベース

* 岡山大学学術研究院社会文化科学学域特任助教

2. 1. 1. 日本語を読むための語彙データベース

「日本語を読むための語彙データベース」(VDRJとも、以下「V表」とよぶ)は、読解に必要な語彙力や、読解テキストの語彙の難易度を測定するために開発された、「現代日本語書き言葉均衡コーパス (BCCWJ)」所収の書籍から約2,800万語、および「Yahoo知恵袋」から約500万語を使用頻度順に整理した語彙データベースである¹。異なり語数計141,949語を収録している。使用頻度を反映した「書きことば重要度ランク」のほか、日本語非母語話者の日本語学習のための「旧日本語能力試験出題基準レベル」との対応関係、人文・芸術分野、社会科学分野、理学・工学分野、生物・医学分野における「領域特徴度レベル」といった豊富な語彙情報が含まれている。これまでV表を活用する研究として、沖本 (2020)、田島ほか (2016)、畠山 (2022)、佐藤 (2020)、梁ほか (2024)などと、留学生および日本人学生の語彙力測定に関する研究が数多く見られる。

このデータベースは、日本語非母語話者向けの日本語教育においても、日本語母語話者の国語教育においても重要な語彙データベースであると考え、本研究のデータベース統合の対象の一つに選定した。語彙データベース統合作業では、松下言語学習ラボHPにて公開されているVer. 1.1を使用する。このデータベースは「重要度順語彙データベース (Top 60894)」「想定既知語彙データベース フィラー、記号その他、固有名詞」「使用範囲狭小語彙データベース」の3ファイルからなり、これらを一つのファイルにまとめてから統合作業を行う。

2. 1. 2. 教科書コーパス語彙表

「教科書コーパス語彙表」(以下「K表」とよぶ)は、2005年度に使用された小学校・中学校・高等学校の全学年・全教科で使用された教科書をテキスト構造化した「教科書コーパス」に出現した語彙を網羅したデータベースである²。異なり語数計50,329語を収録している。「初出学年」のほか、どの学年のどの教科によく出現するかが読み取れる「特徴度」といった語彙情報が含まれている。日本語母語話者、非母語話者の語彙使用状況の測定に用いられる研究として、水谷ほか (2018)、松原・加藤 (2022)、池田 (2023)などが挙げられ、K表を基に開発した語彙力測定システム「ごいす〜」もある。

このデータベースは、特に語彙の使用分野や、日本語母語話者にとっての難易度を知るために必要なものとして、本研究のデータベース統合の対象の一つに選定した。語彙データベース統合作業では、2011年に国立国語研究所が公開した「特定領域研究『日本語コーパス』言語政策班最終成果CD-ROM」所収の「教科書コーパス語彙表」(Ver 1.0)を使用する。

¹ 松下 (2010) を参照。

² 田中・近藤 (2011) を参照。

2. 1. 3. 日本語教育語彙表

「日本語教育語彙表」(以下「N表」とよぶ)は、バイリンガルの日本語学習辞書の開発を可能にするために、日本語教科書および「現代日本語書き言葉均衡コーパス (BCCWJ)」の語彙調査を基に作成されたデータベースである³。計17,920語を収録している。ダウンロード版では、各見出し語に「語彙の難易度」という語彙情報が付いており、日本語非母語話者にとっての語彙難易度を「初級前半」「初級後半」「中級前半」「中級後半」「上級前半」「上級後半」の6段階に整理している。オンライン版では、さらに「アクセント」「意味特性」「旧日本語能力試験出題基準レベル」「語義と用例」といった語彙情報が確認できる。語彙難易度の判定について、旧日本語能力試験出題基準による研究が多かったが、2010年に新試験 (JLPT) への移行に伴い出題基準が非公開となつて以降は、N表を語彙難易度判定の基準とする研究が増えている。近年の研究として、松本 (2020)、山元 (2023) などが挙げられる。

このデータベースは、旧日本語能力試験出題基準と異なる視点で、日本語非母語話者にとっての語彙難易度を知るために必要なものであるため、本研究のデータベース統合の対象の一つに選定した。語彙データベース統合作業では、「日本語教育語彙表」プロジェクト紹介ページにて公開中のダウンロード版 (Ver 2.8.3) を使用する。

2. 1. 4. 日中対照漢字語データベース

「日中対照漢字語データベース」(JKVCとも、以下「J表」とよぶ)は、中国語母語話者のための日本語教育のために作成された、日本語漢字語彙と中国語との意味対応関係、字体の違い、意味のズレなどが調べられるデータベースである⁴。計10,053語を収録している。中国語と日本語の間に意味のズレがある見出し語については、「日本語独自の意味の類推可能性」という語彙情報で中国語母語話者からの意味推測のしやすさを示している。日中同形語研究について、文化庁 (1978) を用いる研究が多かったが、それから半世紀近くの時間が経って新しく作られたJ表には大きな意味があり、これを活かした研究として曹 (2023) などが挙げられる。

このデータベースは、日本語教育だけでなく中国語との対応関係を知ることができ、今後は言語横断的研究を進めるために一石を投じるデータベースであると考えため、本研究のデータベース統合の対象の一つに選定した。語彙データベース統合作業では、松下言語学習ラボHPにて公開中のVersion 2.00を使用する。

³ Yurikoほか (2012) を参照。

⁴ 松下ほか (2020) を参照。

2. 2. 語彙の収録対象

本統合型語彙データベースは、外国人を対象とする日本語教育、日本人を対象とする国語教育、並びに専門教育等への応用を見据えて、読みを持つ語彙のみを収録対象とし、読みを持たない記号類等は対象外とする。例えば、「々」は繰り返しを示す記号だが、「国々（グニ）」「人々（ビト）」のように使用される環境によって読みが不定であるため、本データベースの収録対象外とする。一方、「株」は「株式会社」を意味する記号で「カブシキガイシャ」と読まれる安定性を持つため、記号類ではあるものの、読みをもつものとして、本データベースの収録対象とする。

なお、本研究開発では「一語一音」を原則とする。複数の読みをもつ語彙項目については、読みのパターン分だけ、当該語彙項目を分割する（3. 1. 5節、3. 4. 2節を参照）。

3. 収録語彙の考察と処理

3. 1. 日本語を読むための語彙データベース（V表）

3. 1. 1. 読み情報を欠く見出し語の削除

V表では、見出し語の読み⁵として「標準的読み方（カタカナ）」および「語彙素読み」の2欄を設け、カタカナによる読みを示している。V表凡例によれば、前者は天野・近藤（2003）、後者はUniDicによる。この2欄ともに以下a～cのいずれかに該当する場合、読みの確認ができないものとして、2. 2節で述べたように本統合型データベースの収録対象から取り除く。

- a. 空欄
- b. 「0」表記
- c. 「#N/A」表記

但し、他データベースや辞書等から容易に読みが確認できるものについては、その読みを追加することで収録対象とする。

〈表1〉V表：読みを追加する見出し語（一部）

ID	見出し語彙素	語彙素読み	修正前	修正後
			標準的読み方(カタカナ)	標準的読み方(カタカナ)

⁵ 本稿では、実際の発音ではなく、ひらがなやカタカナでどう表記するか綴りのことを「読み」という。例えば「こんにちは」の場合、「コンニチワ」ではなく「コンニチハ」を「読み」とする。以下同様。

77325	html	0	0	エイチティーエムエル ⁶
101303	dll	0	0	ディーエルエル
53305	amazon	0	0	アマゾン
101237	ipod	0	0	アイポッド

なお、「標準的読み方（カタカナ）」と「語彙素読み」が異なる場合、前記a～cの場合を除き、基本的に「標準的読み方（カタカナ）」の示す読みを採用する。V表を土台にして開発された日中対照漢字語データベース（J表）では、見出し語の読みを「標準的読み方（カタカナ）」のみ（「語彙素読み」欄なし）で示しているからである。

3. 1. 2. 注記等とみられる文字の削除

一例として以下に掲げるように、「見出し語彙素」欄に「𪛗」「𪛘」「𪛙」「𪛚」「𪛛」「𪛜」「𪛝」「𪛞」といった文字を使用した注記やマークの代用が見られる。新たに「注」の欄を設けて、必要に応じて注記情報を移動させることで、「見出し語彙素」欄の表記を修正する。

〈表2〉V表：注記等とみられる文字の処理（一部）

ID	標準的読み方 （カタカナ）	修正前	修正後	
		見出し語彙素	見出し語彙素	注
1183	バス	バス𪛗bus	バス	bus
6107	マッチ	マッチ𪛗match 𪛘合致	マッチ	match/ 合致
18900	タラ	タラ𪛗外𪛙tara	タラ	外/tara
101934	コンパネ	コンパネ𪛗concrete 𪛚panel	コンパネ	concrete-panel
50603	アイル	アイル𪛗i 𪛛ll	アイル	I'll（?）
64056	タラ	タラ𪛗町𪛜郡	太良	町／郡
14580	ハウジョウ	ハウジョウ𪛗町𪛜郡𪛝市	ハウジョウ	町／郡／市
53923	サンピエトロ	サン𪛞ピエトロ	サン・ピエトロ	

3. 1. 3. アルファベット等による表記の統一

「見出し語彙素」欄にアルファベット等がある場合、すべて半角にする。なお、特に表記に決まりがない場合、頭字語（一連の単語の最初の文字による略語）は大文字表記にし、それ以外は小文

⁶ 他辞典類では、例えば『広辞苑（第七版）』では「エッチティーエムエル」となっている。一方、教科書コーパス語彙表（K表）では「エイチティーエムエル」とあるため、データベース間の統合性を考慮し、「エイチティーエムエル」という読みを採用した。

字表記にする。但し、製品名などは公式HP掲載の表記を基準とする。

〈表3〉V表：アルファベット等による表記の処理（一部）

ID	修正前	修正後
	見出し語彙素	見出し語彙素
25932	ime	IME
112513	geforce	GeForce

Input Method Editorの略
公式HP参照

3. 1. 4. 地名の漢字表記の復元

V表では、市町村名をはじめとする地名（固有名詞）は基本的にカタカナで表記されている。そのうち、漢字表記の揺れがほぼ見られない地名の表記を漢字に改める。

〈表4〉V表：地名の表記の処理（一部）

ID	書字形（例）	修正前	修正後
		見出し語彙素	見出し語彙素
287	東京	トウキョウ	東京
4736	静岡	シズオカ	静岡
3605	上海	シャンハイ	上海
9559	大連	ダイレン	大連

3. 1. 5. 複数の読みを持つ見出し語の分離

「標準的読み方（カタカナ）」の欄では、記号等の使用を避ける。したがって、下記の見出し語に対しては、2語に分割する処理を行う。

〈表5〉V表：複数の読みを持つ見出し語の処理

ID	修正前		修正後	
	見出し語彙素	標準的読み方(カタカナ)	見出し語彙素	標準的読み方(カタカナ)
114	他	タ／ホカ	他	タ
			他	ホカ
8842	守	(ノ) カミ	守	カミ
			守	ノカミ
27309	alt	オルト／エーエルティー	alt	オルト
			alt	エーエルティー

3. 1. 6. 読みの誤りの修正

V表「標準的読み方（カタカナ）」の欄には以下のような誤りがあったため、これを修正する。

〈表6〉V表：読みの誤りの修正

ID	見出し語彙素	修正前	修正後
		標準的読み方（カタカナ）	標準的読み方（カタカナ）
37	来る	コイ	クル
679	では	デワ	デハ ⁷
937	付き	ツク	ツキ
1156	あっ	アー	アッ
3539	放し	ハナス	ハナシ
3797	えっ	エー	エッ
4562	ごと	コト	ゴト
4565	有り	アル	アリ
4597	房	ふさ	フサ
4615	合い	アウ	アイ
5044	勤め	ツトメル	ツトメ
5054	持ち	モツ	モチ
5082	帰す	キスル	カエス ⁸
5296	連れ	ツレル	ツレ
5313	ミエ ⁹	サンジュウ	ミエ
5359	書き	カク	カキ
5563	なあ	ナ	ナア
5590	勤め	ススメル	ススメ
5971	掛かり	カカル	ガカリ ¹⁰
6043	届け	トドケル	トドケ

⁷ 綴りではなく実際の読みを表記しようとして「デワ」とした可能性もあるが、類例を見る限り、ID-313：或いは（アルイハ）、ID-7214：今日は（コンニチハ）などのようにワと読まれるはずの見出し語もすべてハと表記されている。したがって、この「デワ」という表記は不適切なものとして、「デハ」に改める。

⁸ 「帰す」は「キス」とも読めるが、「語種」欄は「和」（和語）、かつ「語彙素読み」は「カエス」とあるため、「カエス」と読まれるものと考えられる。

⁹ 「品詞」欄には「名詞-固有名詞-地名-一般」、かつ「書字形（例）」欄には「三重」とあるため、地名の「三重」（ミエ）を指すものと思われる。

¹⁰ 「品詞」欄は「接尾辞-名詞的-一般」となっており、かつ「語彙素読み」の欄には「ガカリ」とあるため、「ガカリ」と読まれるものと考えられる。

6187	好く	ヨイ	スク
6495	断わり	コトワル	コトワリ
7037	梁	ハル	ハリ
7060	包み	ツツム	ツツミ
7829	置き	オク	オキ
7867	焼け	ヤケル	ヤケ
8019	入れ	イレル	イレ
8136	遅く	オソイ	オソク
8383	図形	ガンメン	ズケイ
8384	顔面	ズケイ	ガンメン
8651	漬け	ツケル	ツケ
9470	こら	コレハ	コラ
9883	樟	タリ	タル
10402	容れる	イル	イレル
11139	聞き	キク	キキ
11818	温か	シラギ	アタタカ
11819	奮闘	ノミ	フントウ
11820	新羅	ソウジテ	シラギ
11823	縮む	フントウ	チヂム
11824	総じて	チヂム	ソウジテ
12219	ぐるみ	クルミ	グルミ
12313	はあ	ハー	ハア
12576	察し	サッスル	サッシ
12723	どき	トキ	ドキ
12801	分け	ワケル	ワケ
13168	ぼうっと	ボオット	ボウット
13480	生ける	イキル	イケル
13988	瓜	ツメ	ウリ
14271	揚げ	アゲル	アゲ
14398	吹き	フク	フキ
14581	果せる	ハタス	オオセル
14614	欠け	カケル	カケ

14664	巻き	マク	マキ
14740	誓い	チカウ	チカイ
15285	こく	コイ	コク
16442	清らか	ヤネウラ	キヨラカ
16443	損得	サラシ	ソントク
16446	晒し	キナガ	サラシ
16448	寄稿	ムシバム	キコウ
16451	惨事	キヨラカ	サンジ
16452	蝕む	ソントク	ムシバム
16797	止まり	トマル	トマリ
16846	クロス	クロース	クロス
17187	飼い	カウ	カイ
17306	触れ	フレル	フレ
18734	織り	オル	オリ
19313	ひー	ヒト	ヒー
19369	幟	ノボル	ノボリ
19739	炒め	イタメル	イタメ
21326	樋	トク	トイ
22301	掘り	ホル	ホリ
22365	ちゅう	ツウ	チュウ
26340	ptsd	ピーティーエスディ	ピーティーエスディー
27900	仏国	ブッコク	フツコク ¹¹

3. 2. 教科書コーパス語彙表 (K表)

3. 2. 1. 促音便を伴う接頭辞の読みの修正

K表で接頭辞に分類されている「突」「絶」は、「語彙素読み」欄では「トッ」「ゼツ」と促音便を起こした形となっている。この二語はV表にも収録があり、「標準的読み方」「語彙素読み」の二欄ともに促音便を起こさない「トッ」「ゼツ」になっている。確かに、「突端 (トッパシ)」「突始め (トッパジメ)」「絶好調 (ゼツコウチョウ)」のように促音便を起こす場合も多く見られるが、「突端」

¹¹ 「仏の国」「仏教の国」を意味する一般名詞「仏国 (ブッコク)」も確かに存在する。しかしV表の見出し語の「品詞」欄には「名詞・固有名詞・地名・国」とあるため、明らかに「フランス」を意味する「仏国 (フツコク)」であると思われる。したがって、この見出し語の読みを「フツコク」に改める。

「突始め」のように同時に後部要素頭部に半濁音化が起きていること、また「絶不調（ゼツフチョウ）」のように促音便が起きない場合もあることを考えれば、V表のように促音便を起こさない「トツ」「ゼツ」という読みにしたほうが妥当だと思われる。したがって、「突」「絶」の「語彙素読み」を「トツ」「ゼツ」に改める。

〈表7〉K表：接頭辞の読みの修正

ID_教科書	語彙素	修正前	修正後
		語彙素読み	語彙素読み
24614	突	トツ	トツ
32258	絶	ゼツ	ゼツ

3. 2. 2. 同表記で読みが異なる見出し語の処理

漢字表記が同じでも、読みが異なれば、基本的に別項にする。

〈表8〉K表：漢字表記が同じ語の処理（一部）

ID_教科書	語彙素	語彙素読み	V表		処理
			標準的読み方(カタカナ)	語彙素読み	
282	銀白	ギンハク	ギンパク	ギンパク	別項目として新規追加
5708	大山祇	オオヤマツミ	#N/A	オオヤマツミ	同一項目に認定

アルファベット表記の見出し語も同じく、読みが異なれば基本的に別項にする。

〈表9〉K表：アルファベット表記が同じで読みが異なる語の処理

ID_教科書	語彙素	語彙素読み	V表		処理
			標準的読み方(カタカナ)	語彙素読み	
75	IEEE	アイイー イーイー	#N/A	アイトリプル イー	別項目として新規追加
130	ICAO	アイシー エーオー	イカオ	イカオ	別項目として新規追加

4480	ARPA	エーアール ピーエー	#N/A	アーパ	別項目として 新規追加
4793	NIC	エヌアイシー	ニック	ニック	別項目として 新規追加
4795	NEDO	エヌイー ディーオー	ネド	ネド	別項目として 新規追加
4858	FD	エフディー	フロッピーディスク	エフディー	別項目として 新規追加
18261	CAM	シーエーエム	#N/A	キャム	別項目として 新規追加
37412	PEG	ピーイージー	#N/A	ペグ	別項目として 新規追加
47112	UNIDO	ユニード	ユニド	ユニド	別項目として 新規追加

ただ、V表、K表ともに意味の記述がなく、語義の特定が困難なケースが多い。中には、多義語と見られる語が同一項目にまとめられていることが見受けられる。例えば、「CD（シーディー）」は、記録媒体の一つであるコンパクトディスクを指すほか、数学では点Cと点Dを繋げた線（辺や弧）を指す用法もある¹²。

3. 2. 3. 活用型の修正

「活用型」欄の内容について、V表との整合性を考慮し、「サ行変格」で「～ずる」で終わるものはすべて「ザ行変格」に改めた。

〈表10〉K表：活用型の修正

ID	見出し語彙素	標準的読み方 （カタカナ）	修正前	修正後
			活用型	活用型
1244	甘んずる	アマンズル	サ行変格	ザ行変格

¹² K表で「CD」の「特徴度_小中高_数学」は258.87と極めて高い水準を示しているが、辺や弧を指す用法であると考えられる。一方、「特徴度_小中高_技術家庭」も50.49あり、これは記録媒体を指す用法かと思われる。以上から、K表は語形を基に各教科の語彙をリストアップしたもので、意味用法による細分類は行われていないと思われる。

4393	詠ずる	エイズル	サ行変格	ザ行変格
5351	応ずる	オウズル	サ行変格	ザ行変格
6392	重んずる	オモンズル	サ行変格	ザ行変格
9308	軽んずる	カロンズル	サ行変格	ザ行変格
9701	感ずる	カンズル	サ行変格	ザ行変格
11345	興ずる	キョウズル	サ行変格	ザ行変格
11966	禁ずる	キンズル	サ行変格	ザ行変格
14113	献ずる	ケンズル	サ行変格	ザ行変格
14114	減ずる	ゲンズル	サ行変格	ザ行変格
14856	高ずる	コウズル	サ行変格	ザ行変格
17170	先んずる	サキンズル	サ行変格	ザ行変格
20974	殉ずる	ジュンズル	サ行変格	ザ行変格
20975	準ずる	ジュンズル	サ行変格	ザ行変格
21406	生ずる	ショウズル	サ行変格	ザ行変格
21409	乗ずる	ジョウズル	サ行変格	ザ行変格
22577	信ずる	シンズル	サ行変格	ザ行変格
26196	諳んずる	ソランズル	サ行変格	ザ行変格
26269	存ずる	ゾンズル	サ行変格	ザ行変格
28100	嘆ずる	タンズル	サ行変格	ザ行変格
28101	断ずる	ダンズル	サ行変格	ザ行変格
29051	長ずる	チョウズル	サ行変格	ザ行変格
29465	通ずる	ツウズル	サ行変格	ザ行変格
31018	転ずる	テンズル	サ行変格	ザ行変格
31562	投ずる	トウズル	サ行変格	ザ行変格
31564	動ずる	ドウズル	サ行変格	ザ行変格
33773	難ずる	ナンズル	サ行変格	ザ行変格
34360	任ずる	ニンズル	サ行変格	ザ行変格
34764	念ずる	ネンズル	サ行変格	ザ行変格
41734	奉ずる	ホウズル	サ行変格	ザ行変格
45177	命ずる	メイズル	サ行変格	ザ行変格
45180	銘ずる	メイズル	サ行変格	ザ行変格
45504	免ずる	メンズル	サ行変格	ザ行変格
46461	安んずる	ヤスンズル	サ行変格	ザ行変格

3. 3. 日本語教育語彙表 (N表)

3. 3. 1. 読みの修正

N表には読みの誤りが散見される。以下の見出し語に対して、読みの修正を行う。

〈表11〉 N表：読みの誤りの修正

No	標準的な表記	修正前	修正後
		読み	読み
4022	気に入る	キニハイル	キニイル
5180	月曜日	ゲツヨウヒ	ゲツヨウビ
10846	旅館	タビヤカタ	リョカン
11002	誕生日	タンジョウヒ	タンジョウビ
11943	出入り口	デイリクチ	デイリグチ

3. 3. 2. 見出し語の削除

明らかに読みの誤りがあると考えられる場合でも、元の読みの特定ができない場合は、当該見出し語を削除する。

〈表12〉 N表：読みの特定ができないための削除

No	標準的な表記	修正前	削除の理由
		読み	
779	どれ	イズレ	「No-13009 どれ(ドレ)」が既に別項に存在するため。
843	一昨日	イチキノウ	「No-948 一昨日 (イッサクジツ)」 「No-2151 おととい (オトトイ)」 が既に別項に存在するため、「オトツイ」などと読まれるものか。 (読みの特定が難しい)
872	一つ	イチツ	「No-15070 一つ (ヒトツ)」 が既に別項に存在するため。
880	一人一人	イチニンヒトリ	「No-15102 一人一人 (ヒトリヒトリ)」 が既に別項に存在するため。
899	一昔	イチムカシ	「No-15087 一昔 (ヒトムカシ)」 が既に別項に存在するため。

7935	十分	ジュウブン	「ジュップン」「ジップン」なのか「じゅうぶん」なのかの特定できないため。
13471	日当たり	ニチアタリ	「No-14781 日当たり（ヒアタリ）」が既に別項に存在するため。
20012	アナウンサー	アナウンサ	「No-337 アナウンサー（アナウンサー）」が既に別項に存在するため。
20024	エスカレーター	エスカレータ	「No-1638 エスカレーター（エスカレーター）」が既に別項に存在するため。
20025	エレベーター	エレベータ	「No-1719 エレベーター（エレベーター）」が既に別項に存在するため。
20058	スピーカー	スピーカ	「No-9175 スピーカー（スピーカー）」が既に別項に存在するため。
20069	ドライバー	ドライバ	「No-12911 ドライバー（ドライバー）」が既に別項に存在するため。

3. 4. 日中対照漢字語データベース（J表）

3. 4. 1. 読みの修正

J表には読みの誤りが見られる。該当箇所はV表「標準的読み方（カタカナ）」欄と一致しており、V表からの誤った読み情報を引用したことに原因があるかとみられる。

〈表13〉J表：読みの誤りの修正

ID	見出し語彙素	修正前	修正後
		標準的読み方（カタカナ）	標準的読み方（カタカナ）
04068	図形	ガンメン	ズケイ
04069	顔面	ズケイ	ガンメン
05557	奮闘	ノミ	フントウ
07562	損得	サラシ	ソントク
07564	寄稿	ムシバム	キコウ
07565	惨事	キヨラカ	サンジ

3. 4. 2. 複数の読みを持つ見出し語の分離

複数の読みをもつ見出し語に対して分離処理を行った。

〈表14〉J表：複数の読みを持つ見出し語の処理

ID	修正前		修正後	
	見出し語彙素	標準的読み方(カタカナ)	見出し語彙素	標準的読み方(カタカナ)
00007	年	ネン／トシ	年	ネン
			年	トシ
00015	月	ゲツ／ツキ	月	ゲツ
			月	ツキ
00021	歳	サイ／トシ	歳	サイ
			歳	トシ
00045	朝	チョウ／アサ	朝	チョウ
			朝	アサ
00049	町	チョウ／マチ	町	チョウ
			町	マチ
00130	角	カク／かど ^{ママ}	角	カク
			角	カド
00198	君	クン／キミ	君	クン
			君	キミ
00252	寺	ジ／テラ	寺	ジ
			寺	テラ
00256	島	トウ／シマ	島	トウ
			島	シマ
00299	鏡	キョウ／カガミ	鏡	キョウ
			鏡	カガミ
00308	軒	ケン／ノキ	軒	ケン
			軒	ノキ
00352	林	リン／ハヤシ	林	リン
			林	ハヤシ
00412	的	テキ／ ^{ママ} 13	的	テキ
			的	マト

¹³ 「語種」には「漢／和」とあるため、和語（訓読み）としての「的」が落ちていると思われる。「的」字の訓読みは「マト」としか考えられないため、これに改めた。

00413	者	シャ／モノ	者	シャ
			者	モノ
00415	他	タ／ホカ	他	タ
			他	ホカ
00433	実	ジツ／み ^{ママ}	実	ジツ
			実	ミ
00484	品	ヒン／シナ	品	ヒン
			品	シナ
00497	際	サイ／きわ ^{ママ}	際	サイ
			際	キワ
00548	局	キョク／つばね ^{ママ}	局	キョク
			局	ツボネ
00564	種	シュ／タネ	種	シュ
			種	タネ
00666	世	ヨ／セイ	世	ヨ
			世	セイ
00840	病	ビョウ／ヤマイ	病	ビョウ
			病	ヤマイ
00955	宮	キュウ／ミヤ	宮	キュウ
			宮	ミヤ
01019	行	ギョウ／こう ^{ママ}	行	ギョウ
			行	コウ
01131	網	モウ／アミ	網	モウ
			網	アミ
01244	刀	トウ／カタナ	刀	トウ
			刀	カタナ
01362	便	ビン／ベン	便	ビン
			便	ベン
01386	源	ゲン／ミナモト	源	ゲン
			源	ミナモト
01629	銭	セン／ゼニ	銭	セン
			銭	ゼニ

01800	祭	サイ／マツリ	祭	サイ
			祭	マツリ
03498	菜	サイ／ナ	菜	サイ
			菜	ナ
03745	造作	ゾウサ／ゾウサク	造作	ゾウサ
			造作	ゾウサク
04813	球	キュウ／タマ	球	キュウ
			球	タマ
05127	俵	ヒョウ／タワラ	俵	ヒョウ
			俵	タワラ

3. 4. 3. 番号の対応ミス

J表では、「留学生用語彙ランク (VDRJ = 松下2011)」欄の番号でV表との対応関係を示しているが、本研究で検証してみたところ、見出し語「ID-000010 二十 (ニジュウ)」より、番号の対応関係がずれている。このため、本研究の語彙同定作業では、「留学生用語彙ランク (VDRJ = 松下2011)」欄の情報を用いずに、独自で語彙の同定作業を行った。

〈表15〉J表：V表との対応関係を示す番号の乱れ (一部)

ID	見出し語彙素	標準的読み方 (カタカナ)	留学生用語彙ランク (VDRJ = 松下2011)	
			誤	正
00010	二十	ニジュウ	93	63.1
00011	時間	ジカン	96	95
00012	万	マン	105	104
00013	車	シャ	109	108
00014	六	ロク	113	112
00015	月 ¹⁴	ゲツ	115	114
		ツキ	115	-

¹⁴ J表では「月」(接尾辞-名詞的-助数詞)の「標準的読み方 (カタカナ)」は「ゲツ／ツキ」となっている。このような場合は「月 (ゲツ)」「月 (ツキ)」の2項に分割してから、他データベースとの語彙同定作業を行う。「月 (ツキ)」(接尾辞-名詞的-助数詞)はV表には収録されていないが、K表 (ID_教科書-29555) およびN表 (No-11618) には収録あり。

3. 5. データベース間の語彙の同定

以上の通りにデータの下処理をした後、データベース間の語彙の同定作業を行う。基本的には、表記、読み、品詞、活用型ともに一致するものを同一項目と認定する。但し、表記が完全に一致しない場合も、読みと品詞が一致していれば、意味用法上の考察をするうえ、同一項目と認められる見出し語は認定する。

〈表16〉統合：表記が一致しない場合の語彙同定（一部）

V表			K表			N表			J表		
ID	見出し語彙素	標準的読み方（カタカナ）	ID_教科書	語彙素	語彙素読み	No	標準的な表記	読み	ID	見出し語彙素	標準的読み方（カタカナ）
2293	祭り	マツリ	43470	祭り	マツリ	16962	祭り	マツリ	01800	祭	マツリ
185	何処	ドコ	32135	何処	ドコ	12707	どこ	ドコ	-		

以上の語彙同定作業で、計132,750語を収録した統合型語彙データベースが完成し、V表、K表、N表、J表の間で語彙情報の相互引用ができるようになった。この統合型語彙データベースのことをJ-CLID¹⁵と称する。

しかし、本研究では対象4表に対する意味解析はほとんど行っておらず、中には誤った見出し語や語彙と認定できない見出し語が含まれていることも否めない。とはいえ、一語一語に対して語彙の認定作業を行うのは、大量の時間を要するうえ、認定者によって判断がまちまちの場合も多いと予想される¹⁶ので、効率的な方法とはいえない。本研究では、改めて独自で語彙の認定作業を行うのではなく、データベース策定基準の多様性を活かし、J-CLIDの収録データベースによるダブルチェックをすることで語彙の認定を行う。つまり、V表、K表、N表、J表のうち2表以上に収録されている見出し語のみを有効語彙として認定する。この方法により、J-CLID:C4より有効語彙44,181語の抽出ができた。

¹⁵ 日本語を基盤とする統合型語彙データベース（Cross-Lexical Integrated Database：Japanese-based version）。今後さらなるデータベース拡充を考慮し、バージョン情報として、データベースの収録数をCに続く数字で示すこととする。2024年8月30日現在の最新バージョンは、V表、K表、N表、J表の4表を収録したJ-CLID:C4 ver1.0である。

¹⁶ 例えば「こんにちは」といった定型表現は語彙と考える人もいれば、語彙ではないと考える人もいる。また、「abs」という見出し語について、専門性によっては「Absolute Value：絶対値」と考える人や、「Anti-lock Brake System：アンチロック・ブレーキングシステム」と考える人、「Asset Backed Securities：資産担保証券」と考える人、あるいは、以上の知識をまったく持たない場合は「abs」という言葉は存在しないと考える人もいるだろう。

4. 今後の課題

本稿では統合型日本語語彙データベース J-CLID:C4 の構築プロセスを述べてきた。しかし、収録データベースである V 表、K 表、N 表、J 表の 4 表はいずれも基本的に語形・読み・品詞などの語彙情報を基に見出し語を分類しており、意味用法による細分類がほとんどされていない。その結果、これら 4 表を統合して作られた J-CLID:C4 も、意味用法を特定できる語彙情報がほとんど入っていない現状である。今後は、同じ見出し語に対して意味用法による細分類をすべく、語彙の意味用法の特定ができるデータベースの取り入れが必要であると考ええる。

また、データベース本体の規模がどれほど大きくても、有効語彙の数が少なければ役に立たない。J-CLID:C4 ではまだ有効語彙の数が少ないが、有効語彙以外の語彙は一概「語彙ではない」としているわけではない¹⁷。有効語彙の正確性と数を向上させるためには、データベースのほか、日本語教科書や参考書、各教科の教科書、専門用語辞典など、幅広い言語資源から語彙を取り入れる必要があると考え、今後の課題としたい。

なお、本稿では紙幅の関係で J-CLID:C4 の構築プロセスに関する記述に留まっているが、日本語教育や語彙力診断システムへの J-CLID の応用については別稿に譲りたい。

<参考文献>

- 天野成昭・近藤公久 (2003) 『日本語の語彙特性』第2期 CD-ROM 版、三省堂
- 池田香葉子 (2023) 「多義動詞派生義の習得を目指した「意味関係学習法」の検討：中国語を母語とする日本語学習者を対象として」『人間文化創成科学論叢』第25巻
- 沖本与子 (2020) 「日本語学習者の助詞・動詞選択における回答時間と誤答率の傾向：5週間のオンライン学習項目の分析を中心に」『言語資源活用ワークショップ2020発表論文集』
- 佐藤尚子・松下達彦・笹尾洋介・田島ますみ・橋本美香 (2020) 「学部入学前日本語予備教育における学術共通語彙知識の獲得」『国際教養学研究』4号
- 曹卓琦 (2023) 「コーパスから見た中国語を母語とする日本語学習者における数量表現使用の問題点」『統計数理研究所共同研究リポート』465号
- 田島ますみ・佐藤尚子・橋本美香・松下達彦・笹尾洋介 (2016) 「日本人大学生の日本語語彙測定を試み」『中央学院大学人間・自然論叢』41号
- 田中牧郎・近藤明日子 (2011) 「教科書コーパス語彙表」『言語政策に役立つ、コーパスを用いた語彙表・漢字表等の作成と活用』国立国語研究所
- 畠山衛 (2022) 「小説やアニメを日本語で理解するために、どの程度の語彙力が必要になるのか」

¹⁷ 例えば「激安 (ゲキヤス)」という言葉は明らかに現代日本語に存在するが、J-CLID:C4 では V 表 (ID-32907) の 1 表のみに収録されているため、有効語彙として認定されていない。

『2022 CAJLE Annual Conference Proceedings』

文化庁 (1978)『中国語と対応する漢語』大蔵省印刷局

松下達彦 (2010)「日本語を読むために必要な語彙とは? : 書籍とインターネットの大規模コーパスに基づく語彙リストの作成」『2010年度日本語教育学会春季大会予稿集』

松下達彦・陳夢夏・王雪竹・陳林柯 (2020)「日中対照漢字語データベースの開発と応用」『日本語教育』177号

松原未和・加藤直樹 (2022)「児童が書いた文章の内容把握を支援するインタフェースの提案と有用性の検証」『情報処理学会研究報告コンピュータと教育』2022-CE-163 5号

松本秀輔 (2020)「中級日本語教科書に取り上げられる副詞について」『同志社大学日本語・日本文化研究』17号

水谷勇介・河原大輔・黒橋禎夫 (2018)「日本語単語の難易度推定の試み」『言語処理学会 第24回年次大会 発表論文集』

山元一晃 (2023)「看護師国家試験における特徴的なカタカナ語の諸相: 留学生支援での活用を目指して」『金城学院大学論集: 人文科学編』20巻1号

梁震・彭悦・笹尾洋介 (2024)「汎用言語モデルは日本語学習者データに基づく語彙難易度を予測できるのか」『言語処理学会第30回年次大会発表論文集』

Yuriko SUNAKAWA, Jae-ho LEE, Mari TAKAHARA (2012) The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries, *Acta Linguistica Asiatica*, Vol.2-2

<参考リンク>

国立国語研究所 (2011)「特定領域研究『日本語コーパス』言語政策班最終成果 CD-ROM」所収「教科書コーパス語彙表 Ver 1.0」(参照日: 2019年7月5日)

<https://repository.ninjal.ac.jp/records/3301>

読解力研究会 (2021) 語彙力測定システム「ごいす〜」

https://dokkai.org/tool/vocab_assess/ (参照日: 2024年8月29日)

日本語学習辞書支援グループ (2015)「日本語教育語彙表」(Ver 2.8.3)

<https://jreadability.net/jev/> (参照日: 2022年4月13日)

松下達彦 (2011)「日本語を読むための語彙データベース (VDRJ) (研究用)」(Ver. 1.1)

<http://www17408ui.sakura.ne.jp/tatsum/database.html> (参照日: 2024年4月3日)

松下達彦・陳夢夏・王雪竹・陳林柯 (2017)「日中対照漢字語データベース」(Version 2.00)

<http://www17408ui.sakura.ne.jp/tatsum/database.html> (参照日: 2024年5月30日)

謝辞 本研究は科研費挑戦的研究(萌芽)23K17595の助成を受けたものである。