

小型国語辞典と現代語コーパスは互いの語彙を どの程度カバーするか

小木曾智信（国立国語研究所）*

荻野真友子（kotoba）、坂倉基（小学館）、岡田純子（国立国語研究所）、原科有里（国立国語研究所）

Mutual Vocabulary Coverage between a Small Japanese Dictionary and Contemporary Corpora

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics (NINJAL))

Mayuko Ogino (kotoba Inc.), Motoi Sakakura (Shogakukan)

Junko Okada (NINJAL), Yuri Harashina (NINJAL)

要旨

本発表では、小型国語辞典の見出し語と現代日本語コーパスの語彙との対応関係を定量的に分析する。対象とする『現代国語例解辞典 第五版』の全見出し語について、短単位列への対応表を作成し、BCCWJ等の現代語コーパスにおける出現状況を調査する。これにより、国語辞典見出し語のうちどの程度がコーパスから実例を取得できるか、またコーパス語彙のうちどれだけが国語辞典見出し語に対応づけられるかを、頻度情報とともに明らかにする。さらに、レジスターの異なるコーパスごとに比較し、小型国語辞典が現代日本語の各レジスターをどのように反映しているかを語彙的特徴とともに検討する。

1. はじめに

日本語コーパスの整備が進み、国語辞典の編纂にあたってコーパスが参考資料として用いられることが増えてきた。国語辞典はコーパスに対する形態論情報付与の重要な参考資料であり、コーパスは国語辞典の用例収集や使用状況を確認するための重要な参考資料である。これまでも個々の見出し語の語義記述等については、国語辞典とコーパスの用例が比較・検討されることはしばしばであった。しかし、国語辞典とコーパスそれぞれの語彙の総体がどのような関係にあるのかを調査した研究は、管見に入る限り見当たらないようである。見出し語のサイズが異なり、認定基準が相互に異なるため、そもそも一方の見出し語が他方にも採られているのかどうかさえ簡単には確定できない。このことが、国語辞典とコーパスの語彙全体を結びつける研究を妨げている。本発表は、小型国語辞典と現代語コーパスを取り上げ、見出し語の対応表を作成して両者を一定の精度で比較することを可能にし、これをもとに相互の語彙がお互いをどれだけカバーしているかを検証するものである。

* togiso@ninjal.ac.jp

2. 調査対象

調査対象の小型国語辞典として、小学館『現代国語例解辞典 第五版』（以下、現国例）（林・松井 2016）を扱った⁽¹⁾。この辞書は、利用者として高校生から社会人までを幅広く対象としており、全体で 71,000 語（「助詞・助動詞」「擬音語」等の付録部分を除いた辞書項目数で 63,000 弱）を収録している。第五版ではコーパスを活用して見出し語や表記欄を整備している。

一方、調査対象としたコーパスは国立国語研究所の『現代日本語書き言葉均衡コーパス 第 1 部』（BCCWJ1）（Maekawa et al. 2014）、『昭和・平成書き言葉コーパス』（SHC）（小木曾ほか 2024）、『日本語日常会話コーパス』（CEJC）（小磯ほか 2023）である。詳細は参考文献に譲るが、BCCWJ1 は現代書き言葉の均衡コーパス：約 12,400 万語、SHC は昭和・平成期の書き言葉の通時コーパス：約 3,800 万語（雑誌・新聞・ベストセラーを 8 年おきに収録）、CEJC は現代の日常会話の書き起こし：約 246 万語（2016 年から 2020 年収録）を収録したものである（語数は短単位基準）。さらに、参考として現時点で利用可能な『現代日本語書き言葉均衡コーパス 第 2 部』BCCWJ2 の一部も対象とした⁽²⁾（山崎ほか 2026）。

3. 国語辞典見出し語とコーパスの対応付け

国語辞典の見出し語とコーパスの見出し語は、見出し語のサイズ、見出し語の認定基準などさまざまな点で違いがある。特に、見出し語のサイズの問題は大きい。国語研のコーパスは、短単位規程によって語の分割基準が定められており（小椋ほか 2011）、短単位に対応した形態素解析用辞書 UniDic（伝ほか 2007）によって見出し語情報が付与されている。短単位は、単位分割の揺れを極力排するため操作的な定義によって作られる短い単位であり、漢語であれば 2 字漢語を原則とする。そのため一般的な国語辞典の見出し語と同じかそれより小さい場合がほとんどである。一方、国語辞典では外形的な規程によらず意味のまとまりを持つ範囲を見出し語として採るため、しばしば長い範囲が見出し語となる。

さらに、そもそも国語辞典では一つの用例を多重に見出し語として採ることが可能であるのに対して、コーパスの短単位では決まった分割方法による見出し語としてしか採ることができないという違いがある。たとえば「重要文化財」という用例があるとき、短単位では規程により「重要」「文化」「財」の 3 語に分けられ（長単位では「重要文化財」の 1 語）、それ以外の見出し語として採られることはない。一方で、国語辞典では「重要文化財」「文化財」「重要」「文化」「財」の 5 つの見出し語が重層的に該当しうる。完全に一致する見出し語が多くの利用者にとって最適であるとしても、構成要素となっている見出し語もまた利用価値のある情報である。

見出し語のサイズ以外にも、同語異語判別の基準、品詞認定基準、異語形のまとめ上げ基準など、さまざまな点で差異がある。この違いは国語辞典とコーパスの間に限ったことではな

⁽¹⁾ 小学館出版局と国立国語研究所の連携協定にもとづき、小学館から『現代国語例解辞典 第五版』の電子データの提供を受けて利用した。

⁽²⁾ 2006 年～2010 年までの書籍：約 2,728 万語、2011 年～2015 年までの書籍コア：約 117 万語、2026 年の SNS（Bluesky）：約 788 万語。現時点では非公開。

く、異なる国語辞典の間にも存在する。こうした違いを含めた厳密な対応付けを行うには、単に見出し語レベルで対応させることはできない。コーパスの用例レベルで意味分類をし直す必要があり、国語辞典ごとに極めて大規模な作業が必要となる。

そこで、本発表では、見出し語認定基準は捨象し見出し語のサイズを対応させることを主眼とした対応表を用意して、コーパスと国語辞典の見出し語を対応づけることとした。

3.1 辞書見出しと UniDic の対応表

上述したとおり、一般に国語辞典の見出し語はコーパスの短単位よりも長い。そこで、見出し語の対応表を作成するにあたり、現国例の見出し語を、短単位の列で表現することを考える。短単位には見出し語のレベルで語彙素 ID が付与されていることから、現国例の見出し語は語彙素 ID 列で表現できることになる。このようなサイズの異なる見出し語の対応表を作成した例として、「分類語彙表」(国立国語研究所 2004) の見出し語を同様に語彙素 ID 列で示した「複数短単位対応版「分類語彙表番号-UniDic」対応表」⁽³⁾(菊池ほか 2025) がある。

表 1 に、現国例の見出し語の一部を語彙素 ID 列で表した例を示す。

表 1 国語辞典見出し語の語彙素 ID 列による表現の例

見出し語	読み	短単位列長	ID_1	ID_2	ID_3
文化	ぶんか	1	33550		
文化財	ぶんかざい	2	33550	14762	
異文化	いぶんか	2	1513	33550	
文化的	ぶんかてき	2	33550	25076	
文化の日	ぶんかのひ	3	33550	28989	31098
重要文化財	じゅうようぶんかざい	3	18245	33550	14762

対象とした現国例の見出し語は辞典本体の 62,872 項目である。この全ての見出し語に対し、次の処理を順に行って、短単位語彙素 ID 列との対応付けを行った。各段階で自動処理の後に人手による修正を加えている。

1. 見出し語・読みによる UniDic 語彙素とのマッチング (短単位一致)
2. 複数短単位対応版「分類語彙表番号-UniDic」対応表を利用したマッチング (複数短単位一致)
3. 以上でマッチしなかった見出し語を UniDic によって短単位解析後、人手修正してマッチング

この作業で現国例の 62,864 項目について語彙素 ID 列の対応表を作成した。現国例の見出し語のうち、最も長い語彙素 ID 列で表現されるものは 6 であった。見出し語の語彙素 ID 列長別の分布を表 2 に示す。83% 以上が短単位と一致するが、約 15% が 2 短単位となる。3 短単位以上は 2% 以下だが、それでも 1,100 語を超える。

なお、現国例において短単位よりも短いために語彙素 ID が付与できなかった見出し語が 8

⁽³⁾ <https://doi.org/10.15084/0002000248>

表 2 現国例見出し語の語彙素 ID 列長別の分布

語彙素 ID 列長	見出し語数	構成比
1	52,355	83.28%
2	9,400	14.95%
3	942	1.50%
4	152	0.24%
5	12	0.02%
6	3	0.00%
合計	62,864	100.00%

例あった。いずれも語の構成要素といえるもので、短単位「一人(ひとり)」「二人(ふたり)」の一部となる「人(り)」のほか、複合動詞の一部となる「突っ/突ん-」「ひん-」、「-かす」「-こくる」「-やか」「-やぐ」である。

4. 見出し語対応表の問題点

前節の手順で作成した見出し語の対応表だが、さまざまなレベルで問題があることがわかっている。なかには国語辞典見出し語とコーパスとの対応付けで生じる本質的な問題で、容易に解決できないものも含まれている。次章以降の議論はこのような制限を前提とした上でのものであることに留意されたい。

4.1 見出し語認定に起因する対応表の制約

見出し語認定基準の違いにより、国語辞典の見出し語と語彙素 ID 列は一対一に対応するとは限らない。[国語辞典見出し語] 対 [語彙素 ID 列] が一対一となる例が圧倒的に多いが、一対多、多対一、多対多のすべての組み合わせが存在する。

4.1.1 多対一で対応するケース

UniDic は一般に異語形を一つの語彙素にまとめ上げるため、国語辞典が異語形を見出し語として採る場合には多対一の関係になる。たとえば、現国例は「ダイヤモンド」「ギヤマン」をそれぞれ見出し語とするが、UniDic 側では一つの語彙素「ダイヤモンド-diamond」(語彙素 ID: 23086) にまとめてその異語形としているため、多対一の関係となる。副詞の「矢張り」「矢っ張り」「矢っ張し」対「矢張り」(語彙素 ID: 38386) など、この種の例は多数ある。

こうした語形のまとめ上げとは異なる原因による場合もある。たとえば、現国例は「空手(からて)」として「てぶら」の意と拳法の一種の 2 つを立てるが、UniDic 側の対応する語彙素は 1 つしかない。同様に、現国例は「建て前」として「本音」の対義語と「上棟式」の意の 2 つを立てるが、UniDic 側の対応する語彙素は 1 つしかない。いずれも語彙素のレベルで二対一の関係となる例である。こうしたものは UniDic 側で見出し語を分割して立項する必要性を感じさせるが、その際には単に見出し語を追加するだけでなくコーパスの用例をそれぞれに割り振る必要が生じることになる。

4.1.2 一対多で対応するケース

逆に、UniDic 側が複数になる場合もある。たとえば現国例で「ダイヤ」は一つの見出し語となっているが、UniDic 側ではこれを「ダイヤ-diamond」「ダイヤ-diagram」の 2 つに区別する。同様に、現国例では、動詞「はかる」（計る・量る・測る・図る・謀る・諮る）は一つの見出し語となっているが、UniDic 側ではこれを「図る」「測る」「諮る」の 3 つに区別する。

品詞認定基準の違いで UniDic 側が多となる場合もある。例えば現国例で「円（えん）」は一つの見出し語だが、UniDic 側ではこれを普通名詞・助数詞の 2 つに区別する。同様に、現国例は「氏（し）」や「方（かた）」が一つの見出し語となっているが、UniDic 側ではこれを普通名詞と接尾辞の 2 つに区別する。

4.1.3 対応づけの課題

多対多の対応は多くないが、助詞「から」や「余（よ）」などは個々の見出し語の中身を見て対応をとる必要がある。以上のように一対一で対応しない場合、仕組み上は、関係データベースを用いて多対多の対応関係を作ることができるが、今回の対応表では、現国例の見出し語に対して便宜的に代表的な語彙素 ID を 1 つだけ採用して、UniDic 側が多となる対応を避けた。国語辞典側の 1 つの見出し語にどこまでの語彙素を対応させるかは、語義に踏み込んだ解釈が不可欠で容易ではないからである。

また、UniDic 側が一となるもののうち語形レベルでも区別できない場合には、コーパスの用例を分類し直さない限り、国語辞典見出し語に対してコーパス用例を正確に対応させることはできない。典型例を検討すればよい国語辞典とは違い、コーパスの全例を分類する立場からは意味ごとの区別は困難な場合が多い。

4.2 マッチングのエラー

このほか、データ集計前にマッチングの処理の誤りを修正しきれなかったものがある。たとえば現国例の「スチール」は steel, steal, still に対応する 3 見出しがあり、UniDic 側もこれに対応した 3 語（及び人名 1 語）があるが、マッチング処理において原語綴りで適切に割り当てることができず、3 語とも「スチール-steel」（語彙素 ID:19242）に割り当てられたまま修正されていない。「コード」code, cord, chord も同様で「コード-chord」（語彙素 ID:82919）が割り当てられている。こうした誤りが複数の同語形の外来語で生じている。同様に、現国例の「投げやり」には槍の一種と「心をこめない」意の 2 つの見出し語があり、UniDic 側もこれを「投げ槍」「投げ遣り」に区別するが、どちらも「投げ槍」（語彙素 ID:62236）に割り当てられている例がある。これらはいずれも単純な見逃しであり、今後修正する予定である。

5. 現国例見出し語のコーパス用例頻度

第 3 節で作成した現国例見出し語の語彙素 ID 列対応表をもとに、SQL 文によって語彙素 ID の連続を検索・集計し、現国例見出し語に 4 つのコーパス（BCCWJ1,BCCWJ2,SHC,CEJC）での頻度を付与した。頻度付きの表（以下「頻度付き見出し語対応表」）の冒頭部分（読み順）を表 3 に示す（ID、読み、語彙素 ID_5-6 を省略した）。

なお、この表の頻度を見るときには次の点に注意する必要がある。第一に、見出し語の長さ

表3 コーパス頻度付きの現国例見出し語の語彙素 ID 列対応表

見出し語	短単位列長	ID_1	ID_2	ID_3	ID_4	BCCWJ1	BCCWJ2	SHC	CEJC
あ	1	772				9833	6189	1670	20165
ああ	1	68				1742	416	1141	399
ああ	1	67				11947	6064	2380	20509
ああいう	2	68	1571			1210	258	786	331
アーカイブ	1	81276				81	388	10	1
アーガイル	1	81278				37	10	1	0
アーク灯	2	81410	25869			4	0	9	0
アーケード	1	1452				208	72	30	5
アース	1	1455				411	172	35	5
アーチ	1	1456				566	156	49	1
アーチェリー	1	1457				47	8	13	0
アーティスト	1	1458				1227	314	66	26
アーティチョーク	1	1459				28	6	1	6
アート	1	1462				2065	2915	224	46
アート紙	2	1462	14947			10	2	0	0
アートディレクター	2	1462	25571			27	11	4	0
アーバン	1	82449				120	15	9	0
アーベント	1	46337				7	2	5	0
アーミー	1	82450				79	17	9	0
アーム	1	46338				375	182	37	0
アーメン	1	46339				96	17	10	0
アーモンド	1	46340				334	152	16	15
アール	1	255096				576	602	26	0
Rh 式血液型	4	93093	15099	11178	8072	5	1	0	0
RNA	1	120216				361	142	144	0
アールデコ	2	1462	43253			18	5	3	0
アールヌーボー	2	1462	46341			6	9	0	1
RV	1	1495				98	35	9	0
相	1	78				41	3	1	0
愛	1	76				7267	3102	2230	21
藍	1	80				462	106	53	2
アイ	1	82453				1216	335	77	15
藹藹	1	46344				120	26	35	1
相合い傘	2	83	6344			7	1	5	0
アイアン	1	46345				245	71	34	40
愛育	1	88				27	6	7	0

が異なるため、コーパスの同じ出現例が、国語辞典見出し語として多重に認定される場合がある。また、前節でみた問題点により、異語形が見出し語となっている場合に語彙素レベルで多重にカウントされる。したがって、この表の集計結果から単純にコーパス全体の何%が辞書の見出し語に採られているかを測ることはできない。

5.1 コーパスの解析エラーと修正

前節のように、コーパスに付与されている形態論情報から各見出し語の頻度を取得することができるが、次に示すようにコーパス側の形態論情報が原因となって対応がとれず、頻度が上手く取得できない場合があった。

まず、大規模な現代語コーパスは非コアデータ中に多数の誤りを含んでおり、それを前提と

して利用せざるを得ないため、対応する用例がとれないという問題がある。

今回のマッチング作業を行う中で、コーパスに当然出現すると思われる見出し語の用例数が頻度 0 となり、その原因がコーパスの解析エラーに起因するものである場合が多数見つかった。今回、こうした例では可能な限りコーパスの形態論情報の修正を行って正しい形態論情報に修正したうえで頻度を取得し直した。したがって、今回対象としたコーパスの形態論情報は一般公開版とは異なるものとなっている⁽⁴⁾。

5.2 書き言葉コーパスの読み曖昧性

また、コーパス側で読みが確定できないために課題が残る場合があった。コーパス中に振り仮名がなく、読みを確定できない場合、BCCWJ 等の書き言葉コーパスではどちらかの読みに全例を寄せている。このような場合にはコーパスの情報に従った。たとえば、次の例では、コーパス中の用例は前者のみで、後者の用例が頻度 0 となっている。

- 上側（ウエガワ／ウワガワ）
- 公孫樹（イチョウ／コウソソング）
- 係助詞（カカリジョシ／ケイジョシ）
- 詩心（シシン／シゴコロ）
- 車幅（シャフク／シャハバ）

5.3 記号が介在する見出し語

さらに、表記上、記号が介在するために語彙素 ID 列でコーパスから用例が取得できないケースがあった。例として、「クロイツフェルトヤコブ病」は、「クロイツフェルト・ヤコブ病」の形でコーパスに出現するが、「・」が「補助記号-一般」として解析されるため語彙素 ID 列が直接対応せず頻度ゼロとなっている。同様の例に「アカデミックハラスメント」があり、これもコーパスでは「アカデミック・ハラスメント」の形でのみコーパスに現れている。こうしたものに対応するにはカタカナ連続の短単位境界に補助記号を含むものも許容するように対応表を拡張するなどの方策が必要となるが、今回の調査では対応を見送った。

6. コーパスから実例を取得できる国語辞典見出し語はどれだけあるか

前節の頻度付き見出し語対応表をもとに、コーパスから実例を取得できる国語辞典見出し語はどれだけあるかを調査した。

6.1 コーパス別の用例取得率

現国例の全 62,864 の見出し語について、4 つのコーパスそれぞれでの用例取得率を表 4 に、また、固有名詞を除いた場合の用例取得率を表 5 に示す。現国例のような小型の国語辞典では固有名詞はあまり収録されていないため、両者に大きな差はない。

用例が取得できるかどうかは、コーパスサイズが大きく影響する。今回対象としたコーパスのサイズは表 6 の通りである（記号等を含む）。

コーパス別用例取得率は、コーパスのサイズと明瞭な正の関係にある。コーパスごとにみた

⁽⁴⁾ 将来的に BCCWJ1,2 などの更新版のコーパスを公開する際には、この修正も反映したものとなる予定である。

表 4 コーパス別用例取得率

コーパス	用例あり	用例なし	取得率
BCCWJ1	58,875	3,989	93.7%
BCCWJ2	53,737	9,127	85.5%
SHC	55,750	7,114	88.7%
CEJC	20,230	42,634	32.2%
4 コーパス全体	60,994	1,870	97.0%

表 5 固有名詞を除いたコーパス別用例取得率

コーパス	用例あり	用例なし	取得率
BCCWJ1	58,264	3,893	93.7%
BCCWJ2	53,189	8,968	85.6%
SHC	55,197	6,960	88.8%
CEJC	20,042	42,115	32.2%
4 コーパス全体	60,326	1,831	97.1%

表 6 調査対象のコーパスサイズ

コーパス	サイズ
BCCWJ1	124,060,494
BCCWJ2	36,574,521
SHC	38,565,902
CEJC	2,463,355

サイズの順位と取得率の順位が完全に一致し、サイズの常用対数と取得率との間にも強い相関がある ($r=0.971$)。ただし、取得率はコーパスの規模だけでなく、収録ジャンルやレジスターにも左右される。この点について、次節でコーパスに用例が見られない見出し語の観点から確認する。

6.2 コーパスに用例が見られない見出し語

6.2.1 どのコーパスにも見られない見出し語

どのコーパスにも見られない見出し語は、現国例の全 62,864 の見出し語中に 1,871 語あった。このうち、読み順の冒頭から 50 語を表 7 に示した。先述した「アカデミックハラスメント」(「・」が入るためマッチしなかったもの)を除き、多くの語が現代語で用例が得がたい古風な語や僻語であるといえる。知らない言葉を調べる、という国語辞典の役割に鑑みれば一定の価値を有するものの、先行辞書から継承され、実際の使用実態を失った「辞書残存語」の候補と見做されるものが含まれている可能性がある。

表7 全コーパス未出現の見出し語（先頭 50 語）

見出し語	読み
合い着	あいぎ
間食い	あいだぐい
文色	あいろ
赤ゲット	あかゲット
赤御飯	あかごはん
アカデミックハラスメント	アカデミックハラスメント
秋落ち	あきおち
空き殻	あきがら
秋蚕	あきご
秋高	あきだか
悪才	あくさい
悪報	あくほう
あぐら鼻	あぐらばな
明け荷	あけに
足ついで	あしついで
足まめ	あしまめ
あぜ編み	あぜあみ
汗染みる	あせじみる
あだ情け	あだなさけ
厚様	あつよう
当て事	あてごと
アテンポ	アテンポ
後押さえ	あとおさえ
浴びせ倒し	あびせたおし
脂足	あぶらあし
脂手	あぶらで
雨あい	あまあい
天つ	あまつ
雨間	あまま
網ジュバン	あみジュバン
歩み合い	あゆみあい
あらあらかしこ	あらあらかしこ
粗皮	あらかわ
荒肝	あらぎも
粗ごなし	あらごなし
新所帯	あらじょたい
新身	あらみ
アルペンシュトゥック	アルペンシュトゥック
暗主	あんしゅ
アントニム	アントニム
イージーペイメント	イージーペイメント
言い破る	いいやぶる
家居	いえい
いかり草	いかりそう
息休め	いきやすめ
遺勲	いくん
居こぼれる	いこぼれる
居敷当て	いしきあて
石投	いしなぎ
遺沢	いたく

6.2.2 SHC にだけ見られる語

SHC にのみ見られる語としては約 1000 語あったが、SHC は全てが非コアデータであり、誤解析を多く含んでいるため精査が必要である。頻度の高いものとして「ぬか油」「唯物弁証法」「肅学」「一弦琴」「農科」「緊切」「ユータナジー」「円価」「簇出」などがあった。この中には SHC に含まれるが BCCWJ には含まれない昭和前期に広く用いられ、その後使われなくなった語が多く含まれていると考えられる。今後、より詳細な調査を行いたい。

6.2.3 CEJC にだけ見られる語

コーパスサイズが小さいこともあり CEJC にだけ見られる語は多くないが、誤解析に起因するものなどを除き、次のような語が見られた。書き言葉コーパスでは読みの曖昧性の問題から音読みの「しりつ」にまとめられてしまっている「市立(いちりつ)」「私立(わたくしりつ)」が、この形で確例として現れている。また、本来「ひとごと」であった(他人事という表記を媒介として成立した)「他人事(たにんごと)」がこの読みで実際に用いられていた。これらが語形を確実なものとして採集できるのは話し言葉コーパスならではの価値であるといえよう。このほか、「骨接ぎ」「寄せ物」などの語があった。

7. 現代語コーパスの短単位見出し語は現国例にどれだけ採録されているか

つづいて、対応表をもとに、コーパス側から見て、コーパスの短単位見出し語がどれだけ辞書に採録されているかを調査する。そのため、複数短単位になる見出し語は考慮せず、見出し語対応表の語彙素 ID 列に有効な ID が一つだけ記載されている辞書見出しを対象とした。4 節で見た多重にカウントされる問題を避けるため、コーパスの頻度は頻度付き見出し語対応表からではなく、コーパスから直接取得した。

集計結果を表 8 に示す。集計にあたって、語彙素 ID が 0 または NULL となる未知語等は異なり語数の分母から除いた。

コーパス側から見たとき、語彙被覆率はコーパスサイズが大きくなるほど低下する。異なり語数ベースの語彙被覆率は 25.5%~57.8% である。全体では 23.0% まで下がる。延べ語数ベースの語彙被覆率は 65.5%~71.8% である。全体で 66.3% となっている。コーパスには辞書に載らない多様な語が現れるため、サイズが大きくなるほど異なり語数ベースの語彙被覆率は低下する。しかし高頻度語も多いため、延べ語数では BCCWJ のサイズとなると安定的で

表 8 短単位に対応する辞書見出しのコーパス語彙被覆率

コーパス	異なり語数			延べ語数		
	対応語数	全語数	被覆率	対応語数	全語数	被覆率
BCCWJ1	48,236	189,069	25.5%	81,073,801	123,710,319	65.5%
BCCWJ2	44,469	122,058	36.4%	23,858,135	36,239,049	65.8%
SHC	46,107	120,330	38.3%	26,510,222	38,520,326	68.8%
CEJC	17,045	29,514	57.8%	1,705,392	2,375,185	71.8%
4 コーパス全体	49,687	215,832	23.0%	133,147,550	200,844,879	66.3%

表 9 固有名詞を除外した短単位に対応する辞書見出しのコーパス語彙被覆率

コーパス	異なり語数			延べ語数		
	対応語数	全語数	被覆率	対応語数	全語数	被覆率
BCCWJ1	47,872	122,481	39.1%	80,806,824	121,063,899	66.7%
BCCWJ2	44,119	88,268	50.0%	23,786,375	35,551,386	66.9%
SHC	45,760	85,004	53.8%	26,355,057	37,361,807	70.5%
CEJC	16,899	24,013	70.4%	1,702,574	2,339,522	72.8%
4 コーパス全体	49,306	139,090	35.4%	132,650,830	196,316,614	67.6%

ある。

つづいて、異なり語数に大きな影響を与える固有名詞を除いた集計結果を表 9 に示す。

固有名詞を除くと、異なり語数の語彙被覆率は大きく上がる。異なり語数ベースの語彙被覆率 39.1%～70.4%。全体では 35.4% となっている。延べ語数ベースの語彙被覆率は大きくは変わらず、66.7%～72.8%。全体で 67.6% となっている。

以上の結果から、小型国語辞典の見出し語は、概ね現代語コーパスの延べ語数の 66% 程度をカバーしていることがわかった。また、異なり語数で見たときには、コーパスサイズにより大きな差があるが、1 億語規模のコーパスの 25% 程度、固有名詞を除くと 40% 程度がカバーされることがわかった。ただし、特に異なり語数では認定基準などの条件によってかなりの幅が出ることが考えられる。

8. おわりに

小型国語辞典の見出し語とコーパスの短単位（語彙素 ID 列）との対応表を作成し、国語辞典の見出し語とコーパスの用例をつなぐことで、国語辞典側から見たコーパス用例の取得率と、コーパス側から見た国語辞典による語彙被覆率を確認することができた。

データ整備が半ばであり十全な結果ではないこと、十分な分析が行えなかったこと、などの憾みが残るが、国語辞典の全ての見出し語を日本語コーパスに接続するおそらく初めての試みとして、一定の成果を得たものと考えたい。

今後、現国例対応データの修正を行うとともに、他の国語辞典、特に中型・大型の国語辞典の見出し語への拡張を行っていきたい。また、古語辞典と歴史コーパスの対応付けも行い、『日本国語大辞典 第 3 版』の編集作業に資する通時的なデータへの拡張を行いたい。こうした辞書・語彙研究の基礎となるデータを整えることで、将来の日本語辞書学の基礎を固めていきたいと考えている。

謝 辞

本研究は国立国語研究所のプロジェクト「多様な語彙資源を統合した研究活用基盤の共創」の成果の一部である。

文 献

- 林巨樹・松井栄一 (編) (2016). 『現代国語例解辞典』 第五版 小学館, 東京
- Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.
- 小木曾智信・近藤明日子・高橋 雄太・間淵洋子 (2024). 「『昭和・平成書き言葉コーパス』の設計・構築・公開」 情報処理学会論文誌, 65:2, pp. 278–291.
- 小磯花絵・天谷晴香・居關友里子・白田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉・渡邊友香 (2023). 「『日本語日常会話コーパス』設計と構築」 国立国語研究所論集, 24, pp. 153–168.
- 山崎誠・呉寧真・近藤明日子・落合哉人・小木曾智信 (2026). 「日本語学会 2026 年度春季大会 ワークショップ BCCWJ2 の構築を通してコーパスの 新展開を考える」 日本語学会 2026 年度春季大会予稿集, pp. 199–212.
- 小椋秀樹・小磯花絵・富士池優美・宮内佐夜香・小西光・原裕 (2011). 『『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版 (下)』, 国立国語研究所内部報告書, <https://doi.org/10.15084/00002856>.
- 伝康晴・小木曾智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007). 「コーパス 日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」 日本語科学, 22, pp. 101–123.
- 国立国語研究所編 (2004). 『『分類語彙表』-増補改訂版データベース』, 国立国語研究所資料集 14, <https://clrd.ninjal.ac.jp/goihyo.html>.
- 菊池そのみ・片山久留美・高橋 雄太・小木曾智信 (2025). 「『複数短単位対応版「分類語彙表 番号-UniDic」対応表』の構築と活用」 国立国語研究所論集, 28, pp. 179–195.