

小中高校教科書に収録される語彙の変化

呂 建輝（大阪大学）[†]

Changes in the Vocabulary Included in Elementary and Secondary Education School Textbooks LU Jianhui (The University of Osaka)

要旨

新学習指導要領の施行により、小中高校で使用される教科書も順次に改訂された。これにより、教科書で学習する語彙にはどのような変化があったのだろうか。本発表では、統合型日本語語彙データベース J-CLID を使い、2005 年に出版された教科書と、2020 年以降に出版された教科書に出現した語彙を比較し考察を行う。さらに、形態素ベースと語彙ベースとで収集した語彙にはどのような違いが見られるかについて考える。

1. はじめに

小中高校教科書にはどのような語彙が使用されているかについて、国立国語研究所が 2011 年に公開した「教科書コーパス語彙表」が知られている。これは 2005 年に使用された小中高校教科書のデータ（教科書コーパス）を、UniDic を用いた形態素解析によって構築された語彙表である。

その後、新学習指導要領の実施により、小中高校教科書には大幅な改定が行われ、教科書に収録された語彙にも大きな変化がみられるはずである。しかし、新学習指導要領に準じて作られた教科書を対象とした語彙調査は、管見の限りまだ見られず、教育語彙の変化の実態は不明である。

そこで本研究では、2020 年以降に文部科学省検定を受けた小中高校教科書を対象に、収録語彙を調査し、「教科書コーパス語彙表」との異同を考察する。

2. 調査方法

「教科書コーパス語彙表」では、教科書本文をコーパス化してから、形態素解析により語彙リストを割り出すという手法が用いられている。この手法では、コーパス化に際して膨大な作業量を伴い、かつ形態素解析により収集されたいわゆる「短単位語」は実際に人間が認識する「語彙」との間にずれがある。そこで本研究では、小中高校教科書の本文ではなく、巻末のキーワード一覧や重要語句一覧にリストアップされている語彙を収集し、調査を行うことにした。「教科書コーパス語彙表」と異なる視点での考察により、「短単位語」視点での考察と「語彙」視点での考察の違いを明らかにする狙いもある。

本研究で調査対象とした教科書は〈表 1〉のとおりである。

〈表 1〉本研究で使用した教科書一覧

教科	出版社	検定年	出版年	教科書タイトル
----	-----	-----	-----	---------

[†] [lujianhui.ciee■gmail.com](mailto:lujianhui.ciee@gmail.com)（■の部分はアットマークに変えてください）

小学理科	学校図書	2019	2020	・みんなと学ぶ小学校理科 3 年 ・みんなと学ぶ小学校理科 4 年 ・みんなと学ぶ小学校理科 5 年 ・みんなと学ぶ小学校理科 6 年
小学社会	教育出版	2019	2020	・小学社会 3 ・小学社会 4 ・小学社会 5 ・小学社会 6
小学国語	光村図書	2019	2020	・こくご 二上 たんぽぽ ・こくご 二下 赤とんぼ ・国語 三上 わかば ・国語 三下 あおぞら ・国語 四上 かがやき ・国語 四下 はばたき ・国語 五 銀河 ・国語 六 創造
小学音楽	教育芸術社	2019	2020	・小学生のおんがく 1 ・小学生の音楽 2 ・小学生の音楽 3 ・小学生の音楽 4 ・小学生の音楽 5 ・小学生の音楽 6
小学家庭	開隆堂	2019	2020	・わたしたちの家庭科 5・6
小学数学	日本文教出版	2019	2022	・しょうがく さんすう 1ねん上 ・しょうがく さんすう 1ねん下 ・小学算数 2年上 ・小学算数 2年下 ・小学算数 3年上 ・小学算数 3年下 ・小学算数 4年上 ・小学算数 4年下 ・小学算数 5年上 ・小学算数 5年下 ・小学算数 6年
中学保健体育	大修館書店	2020	2021	・最新 中学校保健体育
中学技術家庭	開隆堂	2020	2021	・家庭・技術 [家庭分野] ・家庭・技術 [技術分野]
中学国語	東京書籍	2020	2021	・新しい国語 1 ・新しい国語 2 ・新しい国語 3
中学数学	日本文教出版	2020	2021	・中学数学 1 ・中学数学 2 ・中学数学 3
中学理科	学校図書	2020	2021	・中学校 科学 1 ・中学校 科学 2 ・中学校 科学 3
中学社会	教育出版	2020	2021	・中学社会 地理 地域にまなぶ

				・ 中学社会 公民 とともに生きる ・ 中学社会 歴史 未来をひらく
中学音楽	教育出版	2020	2021	・ 中学音楽 1 音楽のおくりもの ・ 中学音楽 2・3 上 音楽のおくりもの ・ 中学音楽 2・3 下 音楽のおくりもの
高校社会	山川出版社	2022	2023	・ 詳説日本史 ・ 詳説世界史
	数研出版	2022	2023	・ 倫理 ・ 公共
	帝国書院	2022	2023	・ 新詳地理探究
	東京書籍	2022	2023	・ 政治・経済
高校数学	数研出版	2021	2022	・ 新編 数学Ⅰ ・ 新編 数学Ⅱ ・ 新編 数学 A
		2022	2023	・ 新編 数学Ⅲ ・ 新編 数学 B ・ 新編 数学 C
高校保健体育	大修館書店	2021	2022	・ 新高等保健体育
高校家庭	実教出版	2021	2022	・ 家庭総合
高校情報	日本文教出版	2021	2022	・ 情報Ⅰ
		2022	2023	・ 情報Ⅱ

なお、教科書より語彙を収集する際、語の読みと意味を逐一確認しており、読みまたは意味の異なる語は別項目としている（例：「分子_(分数の上側にあたる数字)」「分子_(いくつかの原子が結び付いた粒子のこと)」など）。これでは、意味の違いを考慮しない形態素解析辞書 UniDic は使えないので、本研究で語彙の判定に用いたのは J-CLID の語彙体系である。

J-CLID は、筆者の研究チームが開発した統合型日本語語彙データベースである。複数の語彙データベースや市販図書に収録された語彙を「一語一音」「一語一義」の原則に基づき一表に統合したものである。2025 年 5 月より公開中の J-CLID:C7 バージョンには「教科書コーパス語彙表」が収録されているため、2020 年以降の小中高校教科書の語彙表を作成した後、J-CLID の語彙体系に従い語彙同定作業を行うことで、二表間の比較分析が可能である。

以下では、「教科書コーパス語彙表」を A 表とよび、本研究で新たに作成した 2020 年以降の小中高校教科書の語彙表のことを B 表とよぶ。

3. 調査結果と考察

3.1 異なり語数

本研究で 2020 年以降の小中高校教科書を調査した結果、計 15,354 語が集められ、B 表を作成することができた。対して A 表は 50,984 語からなり、語彙数が 1/3 以下に減っていることが分かる。この違いが生じた理由は、A 表は教科書本文から語彙を収集しているのに対して、B 表は巻末の語彙一覧から収集していることにある。巻末の語彙一覧は通常、当該教科書において特に学習者に覚えてほしい言語知識をリスト化したもので、専門性の高い言葉がほとんどであると考えられる。一方、教科書本文から語彙を拾う際には、学術共通語彙（「ある」「いる」「言う」「なる」など）や、日常語彙（「熊」「ライオン」「はさみ」など）も語彙として収集される。このように考えると、もし小中高校教科書の学習語彙の増減が特段に著しいといった事情がなければ、学術共通語彙や日常語彙が教科書本文を占める割合は約 2/3 に及ぶことになる。

3.2 収録語彙の違い

語彙の認定基準について A 表 (UniDic) と B 表 (J-CLID) は異なるものの、それでも A 表に収録された 50,984 語 (短単位語) のうち、計 3,809 語が B 表にも確認できた。これは B 表全体 (15,354 語) の 24.8% を占め、短単位語単独で語彙を構成する場合は 1/4 程度にとどまることを意味する。

B 表所収語彙の最大の特徴は、そのほとんどが名詞で、B 表全体の 99.4% を占めることである。名詞以外の語は〈表 2〉のとおりである。

〈表 2〉 B 表所収の名詞以外の語

品詞	語数	語 例
イ形容詞 (形容詞)	2	等しい (集合 A と B の要素がすべて一致していること)、 等しい (ベクトルの向きと大きさが同じであること)
ナ形容詞 (形状詞)	7	透明、印象的、客観的、効果的、多面的、抽象的、独断的
接頭辞	1	約～
接尾辞 (一般)	5	～国 (こく)、～以上、～以下、～未満、～里
接尾辞 (助数詞)	52	～円、～m (メートル)、～km (キロメートル)、～kg (キログラム)、～倍、～% (パーセント)、～割、～cm (センチメートル)、～秒、～t (トン)、～銭、～g (グラム)、～合、～mm (ミリメートル)、～L (リットル)、～a (アール)、～ミリリットル、～ヘクタール、～ポンド、～ワット、～ミリグラム、～厘、～ジュール、～kcal (キロカロリー)、～byte (バイト)、～m ² (平方メートル)、～kL (キロリットル)、～光年、～hPa (ヘクトパスカル)、～ルクス、～ミリシーベルト、～μg (マイクログラム)、～m ³ (立方メートル)、～ミリアンペア、～シーベルト、～kWh (キロワット時)、～A (アンペア)、～Ω (オーム)、～dL (デシリットル)、～ニュートン、～パケット、～パスカル、～bit (ビット)、～分間、～km ² (平方キロメートル)、～cm ² (平方センチメートル)、～Hz (ヘルツ)、～V (ボルト)、～m/s (メートル毎秒)、～cm ³ (立方センチメートル)、～° (度)、～メツツ
動詞	26	する、見る、知る、焼く、沸かす、含む、接する、属する、量る、解く、拭く、炊く、茹でる、仕立てる、炒める、吸い取る、研ぐ、掃く、繰り上げる、和える、ほつれる、繰り下げる、こする、叩く、掛ける、割る
副詞	3	順次、たつぷり、ひたひた
数詞	1	千

なお、本研究で作成した B 表は UniDic による形態素解析を行っていないため、特に固有名詞の類においては A 表では収集できなかったり、一般名詞として扱われてしまったりする語も集められた (「世界史序説 (イブン・ハルドゥーンの著作)」 「大学 (四書の一つ)」 など)。

3.3 「短単位」基準に統一しての比較

語彙判定基準が異なるようでは A 表と B 表の収録語彙を同一次元で評価することができない。そこで、B 表所収の語彙に対して、sudachipy (分割モードは SplitMode.A に設定) を使い形態素解析を行い、得られた計 13,186 語の短単位語を B' 表に整理した。あらためて A 表と照合したところ、計 9,282 語 (70.4%) が A 表に確認される結果となった。つまり、少なくとも短単位基準で考える場合、新教科書所収語彙のうち約 3 割が新たに追加さ

れた語彙で、新学習指導要領を反映させたものであるといえる。¹

一方、A 表になく B' 表に新たに出現した語彙として、新しい時代を反映した専門語彙（「iPS」「アプリ」「差分」「ニート」「伏線」「ブルトウス」「アウトリーチ」「ラベリング」など）が目立つ。また、言葉自体は一昔からあるものの、新学習指導要領により重点的に学んでほしいという意図が読み取れる語彙も見られた（「LED」「デバッグ」「臥位」など）。

4. 「語彙」と短単位語との違いについて

本研究では、教科書巻末の語彙一覧より語彙を収集、つまり編者が実際に語彙として認識した語を集めて B 表を作成した。形態素解析により構築された A 表との比較を通して、語彙の量的研究における形態素解析の力不足の一面が露呈した。

4.1 誤解析（読みの違い）

自然言語処理では通常、「形態素解析＞構文解析＞意味解析＞文脈解析」という一連のプロセスの中で、文章における特定の語の意味や読みを推測している。しかし、語彙の研究においては形態素解析のみに頼ることが多く、これでは正しく解析できないことがある。例えば、数学用語で「角の大きさ」「正の角」「負の角」「2 平面をなす角」「第 3 象限の角」の「角」は「カク」とよむものと思われるが、形態素解析ではこれらが全て「カド」になる。また、「母標準偏差」「母比率」「母平均」の「母」は「ボ」とよむものと思われるが、形態素解析ではこれらが全て「ハハ」になる。

4.2 多義語の処理

形態素解析では多義語による細分類はなされておらず、品詞分類もあいまいな一面があり、同一語かどうかの判定が難しい。例えば、B 表には人名の「リスト（ドイツの経済学者）」「ボール（ジョン＝ボールのこと）」がみられるが、A 表には普通名詞の「リスト」「ボール」しか収録されていない。2005 年当時の教科書にはなかった新しい知識である可能性もあるが、誤解析により普通名詞に分類されてしまっている可能性も考えられる。似たような例として作品名の「太陽」「日本人」「マハーバーラタ」なども同様である。

4.3 長単位解析の限界

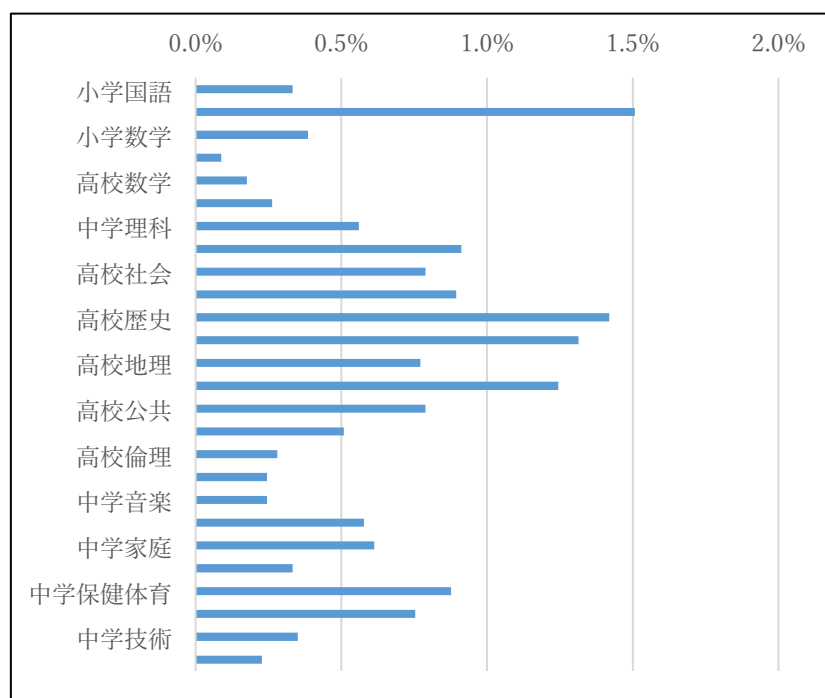
本研究 B 表で収集対象としている「語彙」の判定は、J-CLID に頼らずとも、長単位解析（sudachipy の分割モードを SplitMode.C に設定）を行うことで実現できるという考え方もある。しかし、2026 年現在の長単位解析は、いわゆる臨時一語に対する判定は可能でも、文節を内包する構造の語彙については有効な解析が得られない。例えば「冷戦の終結」と「ベルリンの壁」の二語について、長単位解析では「冷戦 | の | 終結」「ベルリン | の | 壁」と、同様な結果になる。しかし「冷戦の終結」は「冷戦が終結すること」と理解できても、「ベルリンの壁」は東西冷戦のシンボルであり、ただ「ベルリンにある壁」という意味ではない。つまり、「ベルリンの壁」は見かけ上文節からなるようだが、実質一語化しており、現在の長単位解析では一語として認識することができていない。

5. 応用：日本語教科書における語彙の専門性

本研究で作成した B 表は、初・中等教育における専門語彙として位置付けることができる。これを使えば J-CLID 所収の他資料に対して、語彙の専門性を考察することが可能である。例えば、日本国内でよく知られる日本語教科書『みんなの日本語』の学習語彙 5,707 語

¹ 2026 年 8 月現在の UniDic および形態素解析は、A 表が作成された 2011 年と異なるという事情も考慮する必要もある。しかし、少なくとも A 表にも B' 表にも収録のあるこの 9,282 語に限って言えば、当初と現在では変わっていない語彙であるといえる。

について考察したところ、547 語 (9.6%) が小中高校教科書で学ばれる語彙であるということが明らかになった。その内訳は〈図 1〉のとおりである。



〈図 1〉 みんなの日本語（第二版）に収録された語彙の専門性（内訳）

一方、形態素解析によって構築された A 表に照合したところ、5,707 語中 3,869 語 (67.8%) が確認され、本研究で構築した B 表とは全く異なる結果となった。「みんなの日本語」の学習語彙には、短単位語だけでなく「自動販売機」のような長単位語や「早起きは三文の得」のようなことわざも多数収録されている。このように、4 節で述べた事情も踏まえると、少なくとも「語彙」を対象とする量的研究を行う際には、形態素解析のみで結論を出すのには相当の危険性を伴うものとして、十分に注意されたいところである。

6. まとめ

本研究では、新学習指導要領に基づき作成され、2020 年以降に出版された小中高校の教科書を調査し、巻末の語彙一覧に出現した語彙を集めて語彙リスト (B 表) を作成した。これを、2005 年の小中高校教科書に形態素解析を行うことによって構築された「教科書コーパス語彙表」(A 表) との比較分析を行った。その結果、約 3 割の短単位語が新学習指導要領により新たに追加されたものであるということが明らかになった。

異なり語数からみて、B 表は A 表の 1/3 程度の規模にとどまるが、これは、A 表には教科書本文に見られた日常語彙や学術共通語彙も数多く含まれているからだと考えられる。一方、A 表は形態素解析により構築されているため、誤解析や多義語の処理不足、文節を内包する構造の処理の誤りといった問題点も観察される。このことから、A 表は必ずしも語彙情報を正しく反映していない可能性があると考えられ、語彙を対象とする量的研究において注意されたい。

文 献

田中牧郎、相澤正夫、斎藤達哉、棚橋尚子、近藤明日子、河内昭浩、鈴木一史、平山允子

(2011)『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

関連 URL

呂建輝ほか（2025）統合型日本語語彙データベース「J-CLID:C7」 ver 1.0

利用コンテンツ：「教科書コーパス語彙表」「みんなの日本語」

<https://vocabulary-battlers.wixsite.com/japan/j-clid>