「語彙レベル」から見た近代の語彙と現代の語彙
—『太陽コーパス』と『現代日本語書き言葉均衡コーパス』を用いて—

田中 牧郎　（国立国語研究所言語資源研究系）†

Vocabulary Level in Modern and Contemporary Japanese:
Based on Analyses of "Taiyo Corpus" and "Balanced Corpus of Contemporary Written Japanese"

TANAKA Makiro (National Institute for Japanese Language and Linguistics)

１．背景と目的

近代から現代にかけて、日本語の語彙には大きな変化があったが、その変遷の具体的過程は明らかになっていない。こうした研究は、通時的なコーパスを作って記述することによって進めることが望まれよう。国立国語研究所が公開した『太陽コーパス』は、明治時代後期から大正時代を対象とするコーパスで、『現代日本語書き言葉均衡コーパス』は現代を対象とするコーパスである。それぞれが扱う時代の間に昭和時代が欠落しているなど、近代から現代への変遷をとらえるには、この二つのコーパスでは不足だが、明治後期から大正期と現代とを対照することで、近代語から現代語への通時的研究への見通しをつけていくことも可能ではないかと思われる。

語彙の変遷を扱うには、語彙の全体を把握しつつ、変化しない部分と変化する部分とをより分け、変化した部分についてその背景や事情を考察していくことが重要だと思う。そのような研究のためには、語彙を計量的に扱う方法が有効であり、コーパスに付与された形態論情報を用いることが考えられる。『現代日本語書き言葉均衡コーパス』には、形態素解析辞書 UniDic によって形態論情報が付与されているが、『太陽コーパス』にはこれがない。しかし、近時、近代語テキストにも近代語用の UniDic を整備し形態素解析を適用する研究が進んできているので、試験的に『太陽コーパス』に形態論解析を付与したデータを用いて、『現代日本語書き言葉均衡コーパス』のそれと対照してみたい。その際、二つのコーパスの語彙を対照するための枠組みとして、使用頻度に基づいて語彙を階級に分ける「語彙レベル」を用いる。

２．「図書館書籍サブコーパス」と『太陽コーパス』の語彙レベル

『現代日本語書き言葉均衡コーパス』（BCCWJ）は、多様な媒体のサブコーパスから構成されることを特徴としている。筆者らは、特定領域研究「日本語コーパス」言語政策班の研究の一つとして、BCCWJ のサブコーパスのうち６種について語彙調査を行い、「BCCWJ 主要コーパス語彙表」作成し、公開した（田中・近藤 2011）。この語彙表には、６種のサブコーパスごとに、出現した語彙すべての度数、使用率、使用サンプル数、そして語彙レベルの情報が一覧できるようにしたものである。このうち、「語彙レベル」とは、語彙を度数の高いものから順に並べ、上位の語から度数を累積していき、その累積度数が延べ語数の

† mtanaka@ninjal.ac.jp

何パーセントを占めるかという累積使用率（カバー率）によって、５段階に分けたものである。田中・近藤（2011）から、レベルを区画する基準（表１）と、「図書館書籍サブコーパス」（固定長）について五段階に分類した語数（表２）を下に示す。なお、UniDic で付与される品詞情報のうち、助詞・助動詞・記号類・未知語等は対象外としている。

表１　語彙レベルとカバー率

語彙レベル	カバー率（累積使用率）
a	0 - 78%
b	- 88%
c	-94%
d	-97%
e	-100%

表２　図書館書籍における語彙レベルごとの語数

語彙レベル	延べ語数	異なり語数
全体	3,938,696	86,002
a	3,074,655	4,177
b	395,994	6,330
c	242,911	11,595
d	118,642	14,176
e	106,494	49,724

BCCWJ のサブコーパスうち、幅広いジャンルについて書かれていてよく読まれている媒体は何かと言えば、「図書館書籍」であろう。なぜなら、「図書館書籍」は、公共図書館の多くに共通して所蔵されているものからサンプリングされたものだからである。そこで、本稿でも、この図書館書籍の語彙レベルを、現代語の語彙レベルを代表しているものと扱うことにした。

『太陽コーパス』は、1895（明治 28）年から 1928（昭和 3）年まで刊行された総合雑誌『太陽』を対象とするコーパスで、1895（明治 28）年、1901（明治 34）年、1909（明治 42）年、1917（大正 6）年、1925（大正 14）年の５年分の全文がおさめられている。一資料のみが対象だが、この雑誌が当時よく読まれ、幅広い層の著者が広範なジャンルの記事を書いていることから、当時の書き言葉をある程度代表できるものと考えられる。

表３　『太陽コーパス』における語彙レベルごとの語数

語彙レベル	延べ語数	異なり語数
全体	5,384,879	74,089
a	4,199,256	3,476
b	539,920	4,835
c	326,250	8,834
d	161,045	10,177
e	158,408	47,267

この『太陽コーパス』についても、上記と同じカバー率の基準で語彙レベルに分け各語にレベル情報を付与した。その結果を示すと、表３のようになる。表２と表３を比べると、延べ語数は『太陽コーパス』の方がかなり多いが、異なり語数は「図書館書籍」（固定長）の方が多い。二つのコーパスの語彙のありようは異なっていることが見て取れる。

3. 近代の語彙と現代の語彙の対照

3.1　近代は基本的レベルで現代は周辺的レベルの語彙

語彙レベルによって、近代の語彙と現代の語彙を対照した先行研究に、近藤・小木曽（2009）があり、語種構成比率など全体的観点から考察が行われている。本稿では、個別の語にまで焦点を当ててみたい。とくに、近代と現代とで語彙レベルに非常に大きな変動があるものに注目したい。

まず、『太陽コーパス』ではレベル a となっていて、近代では最も基本的なレベルの語彙でありながら、BCCWJ の「図書館書籍」（固定長）では、最も周辺的なレベル e のものを見ていこう[1]。表 4 は、品詞と語種によって分類して、そのすべて（固有名詞を除く）を示したものである。品詞は UniDic の品詞情報をもとに四種に統合した。表４を見ると、名詞-一般の漢語が最も多く、動詞・形容詞・形状詞・副詞には和語や混種語も多くなっている。

表４　『太陽コーパス』でレベル a、「図書館書籍サブコーパス」でレベル e の語彙

品詞	和語	漢語	外来語	混種語
名詞-一般	謂（いい）、魚（うお）、件（くだん）、差し支え、灯し火、一つ	医、位地、一円、一斑、英仏、各省、気球、旭日、現時、効、公債、工兵、国運、国人、三位、爾後、時々、寺内、授、償金、諸種、正貨、政略、責、智、智識、勅令、体（てい）、逓信、徳義、土人、内国、博文、万人、弊、俸給、本邦、命、約、列国	耶蘇	
名詞-サ変等可能		円満、協商、建議、出入、兌換、騰貴		
動詞・形容詞・形状詞・副詞	豈、如何で、怒（いか）る、え、阿る、如此し、希（こいねが）う、然（さ）り、須らく、宜（むべ）、縦し、悪（わろ）い	爾来、超然、畢竟		合（がっ）する、記する、毫も、失する、製する、着する、徴する、変ずる、弁ずる
接辞		該、口、主		

3.2　近代は周辺的レベルで現代は基本的レベルの語彙

次に、『太陽コーパス』でレベル e で最も周辺的なところにあったもので、「図書館書籍」（固定長）では最も基本的なレベル a にある語彙について、一覧にしてみよう（表 5）。表５では、名詞-一般と名詞-サ変等可能の漢語と外来語が特に多くなっており、名詞-一般や動詞・形容詞・形状詞・副詞では和語も比較的多くなっている。

[1] 『太陽コーパス』でレベル a で、「図書館サブコーパス」で全く出現しない語彙を見ることも必要だが、本研究に用いた現代語用の UniDic と近代語用の UniDic とでは、語彙素や品詞の認定が異なる場合があり、機械処理による語彙の対照だけでは、十分な照合ができない部分が残るため、ここではそのタイプは扱わなかった。

表５　『太陽コーパス』でレベル e、「図書館書籍サブコーパス」でレベル a の語彙

品詞	和語	漢語	外来語	混種語
名詞－一般	生き物、兎、餌、親指、方々、側、切っ掛け、気配、塵（ごみ）、汁、食べ物、玉葱、所、主、鼠、初（はつ）、湖	一連、衛星、映像、円、課題、観点、気温、機構、基地、基盤、業績、芸能、現地、高校、次元、視点、時点、集落、衝撃、世代、体制、土（ど）、濃度、便（びん）、複数、米軍、訳（やく）、要因、理念、路線	エンジン、カー、カード、カメラ、カレー、クラス、クリーム、グループ、ケーキ、コース、コスト、サイド、サラリーマン、シーン、ショー、スクール、スタイル、スピード、スペース、タクシー、チーズ、チャンス、テープ、テーマ、トマト、ドラマ、ニュース、パワー、ビジネス、フィルム、プラン、プロセス、ママ、メンバー、リズム、ルール、レストラン、レベル、ワイン	大勢（おおぜい）、生地（きじ）、地元、職場、夕食
名詞－サ変等可能	幾ら、子育て、そう、生（なま）	移行、依存、運営、活性、規制、強調、結成、研修、検討、構想、構築、個別。志向、取材、出演、出土、制作、正常、駐車、直前、定着、定年、展示、統合、導入、入手、入力、配慮、飛行、普段、優先、予感、冷蔵	アップ、オーケー、カット、コントロール、ショック、スイッチ、スタート、セット、ソフト、チェック、デザイン、テスト、バック、バランス、ブルー、プラス、プレゼント、マイナス、ラン	
動詞・形容詞・形状詞・副詞・感動詞	否（いや）、思い込む、組み合わせる、差し出す、そろそろ、辿り着く、次々、取り組む、入（はい）り込む、放る、恵まれる、さす、よし	一体、公的、直接、不可欠		
接辞	形（なり）、焼き	刊、手、人（じん）、帯、道		

文　献

国立国語研究所（編）（2005）『太陽コーパス―雑誌『太陽』日本語データベース』博文館新社

近藤明日子・小木曽智信（2009）「語種を観点とした近代語と現代語の語彙の比較―形態素解析辞書「近代文語 UniDic」「UniDic」を用いて―」言語処理学会第 15 回年次大会

田中牧郎（2010）「雑誌コーパスでとらえる明治・大正期の漢語の変動」（『国際学術研究集会・漢字漢語研究の新次元 予稿集』）

田中牧郎・近藤明日子（2011）「BCCWJ 主要コーパス語彙表」田中・相澤ほか（2011 所収）

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・漢字表の作成と活用』特定領域「日本語コーパス」言語政策班成果報告書