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 ──

松下達彦（Victoria University of Wellington 大学院生）

1．研究概要

本研究では，書籍およびインターネット・フォーラム（Yahoo 知恵袋）で使用される重要な語彙を，使用頻度と使用範囲(range)・分散度(dispersion)のデータに基づいて数万語のレベルまで明らかにし，語彙資料としての性質を検討した。これは，読解に必要な語彙力や，読解テキストの語彙的難易度を測るための基礎資料を作成する研究である。

表1 BCCWJ TM分類 分野別延べ語数
（現代日本語書き言葉均衡コーパス(BCCWJ)2009年モニター版、書籍(BK)およびインターネットQ&Aフォーラム(OC)）

分野	延べ語数	割合
文芸創作	8251999	25.1%
言語・哲学	2134739	6.5%
歴史・民族	3336818	10.2%
芸術、その他の人文科学	3020917	9.2%
政治・法律	1881012	5.7%
経済・商業	2209107	6.7%
社会・教育、その他の社会科学	2996147	9.1%
科学・技術	1512784	4.6%
生物・医学・生活科学	2251037	6.9%
インターネットQ&Aフォーラム	5224852	15.9%
合計	32819412	100.0%

2．研究方法

国立国語研究所の『現代日本語書き言葉均衡コーパス』（BCCWJ）モニター公開データ（2009 年度版）の書籍約 2800 万語のテキストを，日本十進分類法などを手がかりに 9 の下位ジャンルに分類し，「Yahoo 知恵袋」約 500 万語を加えた 10 のジャンルにして（表 1，「TM 分類」），総使用頻度と下位ジャンルごとの相対使用頻度を計算した。使用範囲が広く，かつ使用頻度の高い語彙を重要度の高い語彙と考え，Nation & Beglar (2007) で採用された語彙リスト作成方法（Nation 2010 近刊予定）を参考に，重要度の高い順に 1000 語ずつまとめたベース語彙リスト（「日本語を読むための“TM語彙リスト”」）を 20000 語レベルまで作成した（表 2）。重要度の指標として，いくつかの係数を比較し，Juilland & Chang-Rodrigues (1964) の U（使用度係数）を，下位コーパスの相対頻度に基づいて計算した値が本研究に最適だと考え，これを採用した。これらのリストに基づき，語種別の構成比やテキストカバー率などを調べた。また，ジャンル別の特徴語を対数尤度比(log-likelihood ratio)を用いて抽出した。

表2 BCCWJ 2009モニター版
（書籍、インタ・ネットQ&Aフォ・ラム）
TM語彙リスト 異なり語数

	語彙リスト	語数
ベース語彙リスト（助詞・助動詞125語を含む。）	bw01	1000
	bw02	1000
	bw03	1000
	bw04	1000
	bw05	1000
	bw06	1000
	bw07	1000
	bw08	1000
	bw09	1000
	bw10	1000
	bw11	1000
	bw12	1000
	bw13	1000
	bw14	1000
	bw15	1000
	bw16	1000
	bw17	1000
	bw18	1000
	bw19	1000
	bw20	1000
想定既知語彙リスト	bw01+ (*)	24
	固有名詞 (**)	30440
	フィラー等	11
	記号等	232
	その他	91243
	総計	141950

*bw01+は「二十」などの数詞の複合語で、「透明な複合語」として構成要素の「二」「十」が既知であれば既知語として扱う

**一部の固有名詞は想定既知語彙とはせず、ベースワードに含む

3．分析結果の例，および考察

A) 語種別の構成比は上位 5000 語レベルまで変化するが，6000 語から 20000 語のレベルでは安定している（表 3）。

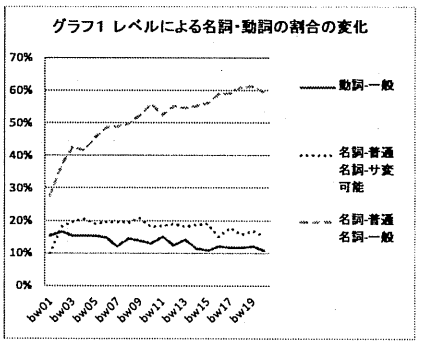
B) 重要度が下がるに連れて，普通名詞の割合は上昇するが，サ変動詞語幹（動名詞）と動詞の割合は 4000/2000 語レベルから緩やかに下降する（グラフ 1）。

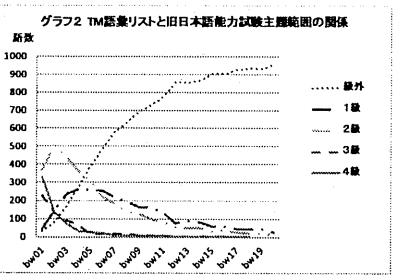
表3 TM語彙リスト 語種別 異なり語彙数

語種	bw01	bw02	bw03	bw04	bw05	bw06	bw07	bw08	bw09	bw10	bw11	bw12	bw13	bw14	bw15	bw16	bw17	bw18	bw19	bw20	総計	割合
和語	529	381	363	335	322	304	311	316	312	338	355	334	327	304	328	350	347	345	348	351	6900	34.5%
漢語	424	531	510	523	495	539	516	541	528	494	486	503	508	538	514	494	510	479	477	482	10092	50.5%
外来語	12	49	81	89	104	95	110	85	111	129	117	121	113	109	116	109	99	124	117	116	2006	10.0%
混種語	19	22	15	14	28	19	17	27	31	28	31	29	44	28	30	28	37	38	36	35	556	2.8%
固有名詞	13	15	26	32	46	34	40	24	13	6	3	4	4	5	4	6	1	3	4	1	284	1.4%
記号、不明等	3	2	5	7	5	9	6	7	5	5	8	9	4	16	8	13	6	11	18	15	162	0.8%
総計	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	20000	100.0%

*bw01～bw20は各1000語からなる、日本語を読むためのベース語彙リストで、bw01が最も基本度が高い。助詞・助動詞計125語を含む。

C) 旧日本語能力試験出題範囲は，読解の語彙は２級から急増するが，bw6 ですでに級外語彙が１級と２級の合計を上回っており，必ずしも日本語を読むために効率的な語彙表とはいえない（グラフ２）。例えば，旧試験の級外語彙でBW1やBW2に入っている語彙には，「自体」「挙げる」「捉える」「表示」「因み（に）」「時点」「多数」「大量」「男女」「そもそも」「層」「初期」「支援」「取得」「当初」「発する」「生み出す」「視線」「応える」等がある。これらの語は新試験ではN3あたりに入る可能性が高い。





これまで教育・研究用に参照された語彙調査には，主に雑誌や新聞に基づくものがあるが，大量の一般書籍やインターネットに基づくものはない。本語彙表はコーパスの延べ語数，語種分布の一般性，語彙の安定度などの点で総合的に優れており，内容的にも例えば大学生に必要な語彙表としては，従来のものより妥当であり，学習・教育用の資料としても，語彙テスト開発の基礎資料としても有用である。

４．今後の課題

本研究で解析に使用したUniDic-MeCabでは語彙素を分析対象としたが，異表記や同形異義語を一まとめにしている場合があり，これを区別できていない。各語の用法を精査して，よりよい語彙表にしたい。また，読解の既知語率を効率的に高められるように，雑誌や新聞等のコーパスも加え，目的に応じた各種の語彙表を作りたい。アカデミック・コーパスを作れば，学問領域共通の語彙や，分野別の専門用語の抽出も可能であろう。作成した語彙表を基にして，テストを開発し，テキストの難易度や読解力などとの関連を調べたい。また，第一言語による想定既知語彙の相違がどう語彙力に影響するかを測定したい。

なお，作成した語彙リストおよび関連資料は，以下のサイトからダウンロードできる。

URL: www.wa.commufa.jp/~tatsum/

引用文献

Juilland, A., & Chang-Rodrigues, E. (1964). Frequency Dictionary of Spanish Words. London: Mouton & Co.

Nation, P. & Beglar, D. (2007) A vocabulary size test. *The Language Teacher* 31(7): 9-13

Nation, I. S. P. (2010, 近刊予定). Making and using word lists. In Nation I. S. P. & Webb, S. *Researching vocabulary.* New York: Heinle Cengage Learning.