「BCCWJ 主要コーパス語彙表」について

近藤 明日子

１．本語彙表の概要

「BCCWJ 主要コーパス語彙表」は、『現代日本語書き言葉均衡コーパス』（BCCWJ）の一部分を構成するサンプル群に出現する語彙に関するデータを収録した電子ファイルである。ファイル名は「BCCWJ.txt」、形式はタブ区切りテキストファイル、文字符号化方式は UTF-16LE（BOM 付き）、改行コードは CR+LF である。

２．語彙調査の概要

本語彙表の作成にあたりおこなった、語彙調査の概要は以下の通りである。

２．１　調査対象テキスト

調査の対象としたテキストは BCCWJ の 2010 年 12 月 9 日版（非公開）のうち、以下の 6 種類である。

(1) 流通実態（図書館）サブコーパスに含まれる書籍の固定長サンプル（LB_FL）、計 10,640 サンプル。
(2) 生産実態（出版）サブコーパスに含まれる書籍の固定長サンプル（PB_FL）、計 10,277 サンプル。
(3) 生産実態（出版）サブコーパスに含まれる雑誌の固定長サンプル（PM_FL）、計 2,439 サンプル。
(4) 生産実態（出版）サブコーパスに含まれる新聞の固定長サンプル（PN_FL）、計 1,489 サンプル。
(5) 非母集団（特定目的）サブコーパスに含まれる Yahoo!知恵袋の可変長サンプル（OC_VL）、計 45,725 サンプル。
(6) 非母集団（特定目的）サブコーパスに含まれる Yahoo!ブログの可変長サンプル（OY_VL）、計 52,680 サンプル。

２．２　テキストの形態素解析

対象テキストの形態素解析結果として、2010 年 12 月 9 日時点で国立国語研究所内のデータベースに集積されているデータを利用した。これは、形態素解析辞書 UniDic[1]（MeCab 版）による解析結果に基づくデータであるが、人手による修正の程度がテキストごとに異なるため、解析精度もまたテキストごとに異なることに留意する必要がある。また、解析結果には解析の誤りが一部含まれ、それに基づく本語彙表にもその誤りが含まれていることにも十分留意する必要がある。

２．３　語の単位

語の単位には UniDic の解析単位である「短単位」を用い、1 短単位を 1 語とした。

[1] http://download.unidic.org/

２．４　同語異語判別

解析結果の同語異語判別は、UniDic により各短単位に付与される属性のうち、「語彙素読み」「語彙素」「語彙素細分類」「語種」「品詞」「活用型」の 6 属性を用い、これらの属性値がすべて一致するものを同語とし、一つの見出し語のもとにまとめた。逆に、6 属性のうち、一つでも属性値が異なれば別語と認定した。表 1 に例としてあげた①～⑦の語では、①と②は「語彙素」、②と③は「活用型」、④と⑤は「語彙素細分類」、⑥と⑦は「品詞」がそれぞれ異なるため別語と認定した。

表１　別語の例

	語彙素読み	語彙素	語彙素細分類	語種	品詞	活用型
①	アラワス	著わす		和	動詞-一般	五段-サ行
②	アラワス	表わす		和	動詞-一般	五段-サ行
③	アラワス	表わす		和	動詞-一般	下一段-サ行
④	オール	オール	all	外	名詞-普通名詞-一般	
⑤	オール	オール	oar	外	名詞-普通名詞-一般	
⑥	アマリ	余り		和	形状詞-一般	
⑦	アマリ	余り		和	副詞	

２．５　収録対象語の選定

本語彙表に収録する見出し語は、調査対象テキスト(1)～(6)のいずれかに 1 回以上出現し、かつ「品詞」属性の大分類が「名詞・代名詞・形状詞・連体詞・副詞・接続詞・感動詞・動詞・形容詞・接頭辞・接尾辞」のいずれかであるものとした。助詞・助動詞・記号類・未知語等は掲載対象外とした。

３．語彙表の構成

本語彙表は、以下の 33 列から構成される。また、先頭行は列名データである。

1	ID_BCCWJ	…本語彙表の掲載された各見出し語に一意に付した番号（語彙素読みの五十音順）
2	ID_全体	…本語彙表および言語政策班の作成した「教科書コーパス語彙表」「学校･社会対照語彙表」に掲載された各見出し語に一意に付した番号（語彙素読みの五十音順）
3	語彙素読み	…UniDic の語彙素読み
4	語彙素	…UniDic の語彙素
5	語彙素細分類	…UniDic の語彙素細分類
6	語種	…UniDic の語種
7	品詞	…UniDic の品詞
8	品詞_大分類	…UniDic の品詞の大分類
9	活用型	…UniDic の活用型
10	度数_LB_FL	…調査対象テキスト(1)での度数
11	度数_PB_FL	…調査対象テキスト(2)での度数
12	度数_PM_FL	…調査対象テキスト(3)での度数
13	度数_PN_FL	…調査対象テキスト(4)での度数
14	度数_OC_VL	…調査対象テキスト(5)での度数
15	度数_OY_VL	…調査対象テキスト(6)での度数
16	使用率_LB_FL	…調査対象テキスト(1)での使用率[2]
17	使用率_PB_FL	…調査対象テキスト(2)での使用率
18	使用率_PM_FL	…調査対象テキスト(3)での使用率
19	使用率_PN_FL	…調査対象テキスト(4)での使用率

[2] 付記 1 参照

20　使用率_OC_VL　　…調査対象テキスト(5)での使用率
21　使用率_OY_VL　　…調査対象テキスト(6)での使用率
22　レベル_LB_FL　　…調査対象テキスト(1)でのレベル[3]
23　レベル_PB_FL　　…調査対象テキスト(2)でのレベル
24　レベル_PM_FL　　…調査対象テキスト(3)でのレベル
25　レベル_PN_FL　　…調査対象テキスト(4)でのレベル
26　レベル_OC_VL　　…調査対象テキスト(5)でのレベル
27　レベル_OY_VL　　…調査対象テキスト(6)でのレベル
28　サンプル数_LB_FL　…調査対象テキスト(1)での出現サンプル数
29　サンプル数_PB_FL　…調査対象テキスト(2)での出現サンプル数
30　サンプル数_PM_FL　…調査対象テキスト(3)での出現サンプル数
31　サンプル数_PN_FL　…調査対象テキスト(4)での出現サンプル数
32　サンプル数_OC_VL　…調査対象テキスト(5)での出現サンプル数
33　サンプル数_OY_VL　…調査対象テキスト(6)での出現サンプル数

４．延べ語数・異なり語数

本語彙表に掲載された見出し語 130,437 語について延べ語数・異なり語数をテキスト・レベル別に示すと表 2 のようになる。

表 2　調査対象テキストの延べ語数・異なり語数

	LB_FL		PB_FL		PM_FL	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
全体	3,938,696	86,002	3,903,395	82,784	896,988	45,900
レベル a	3,074,655	4,177	3,045,639	3,842	700,831	4,336
レベル b	395,994	6,330	391,312	5,609	92,353	5,293
レベル c	242,911	11,595	239,221	10,506	51,085	7,493
レベル d	118,642	14,176	124,601	14,290	37,925	13,984
レベル e	106,494	49,724	102,622	48,537	14,794	14,794

	PN_FL		OC_VL		OY_VL	
	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
全体	624,020	35,727	2,762,864	49,809	6,127,125	76,823
レベル a	486,976	3,420	2,155,871	2,071	4,779,106	3,441
レベル b	63,784	4,045	275,758	2,776	617,945	4,724
レベル c	40,018	6,941	165,957	5,122	372,114	8,406
レベル d	20,607	8,686	83,349	7,062	181,482	10,285
レベル e	12,635	12,635	81,929	32,778	176,478	49,967

付記 1　使用率について

使用率とは、当該テキストの延べ語数に対する当該語の度数の割合を千分率（単位‰）で示したものである。使用率算出に使用した各テキストの延べ語数は本文４を参照のこと。

付記 2　レベルについて

レベルとは、調査対象テキスト(1)～(6)のそれぞれに出現する見出し語について、度数降順の累積使用率

[3] 付記 2 参照

（カバー率）により以下のように a～e の 5 段階に分けたものである（田中、2011）。

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

文献

田中牧郎（2011）「語彙レベルに基づく重要語彙リストの作成―国語政策・国語教育での活用のために―」 本報告書第 2 章第 1 節.