「学校・社会対照語彙表」PDF 版について

近藤 明日子

１．本語彙表の概要

「学校・社会対照語彙表」PDF 版は、特定領域研究「日本語コーパス」言語政策班で作成した「学校・社会対照語彙表」の分類語彙表番号統合版（ファイル名「学校社会対照_統合.txt」）のデータの一部を、紙面上にレイアウトし PDF ファイルとしたものである。

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以下の(1)〜(7)の 7 種類である。

(1) 「教科書コーパス」[1]2010 年 12 月 9 日版（非公開）から、skippedSpan 要素・skippedBlock 要素および type 属性値が newWords または keyWords の supplement 要素を除いた部分。

(2) 『現代日本語書き言葉均衡コーパス』（BCCWJ）2010 年 12 月 9 日版（非公開）のうち、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。

(3) BCCWJ の 2010 年 12 月 9 日版のうち、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。

(4) BCCWJ の 2010 年 12 月 9 日版のうち、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。

(5) BCCWJ の 2010 年 12 月 9 日版のうち、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。

(6) BCCWJ の 2010 年 12 月 9 日版のうち、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。

(7) BCCWJ の 2010 年 12 月 9 日版のうち、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

調査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[2]（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。

---

1 詳細は田中・近藤・平山（2011）を参照。

2 http://download.unidic.org/

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。

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)の中学校・高校教科書部分あるいは調査対象テキスト(2)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．語彙表で使用するカテゴリー

本語彙表で「教科書コーパス」のデータを示す際に用いたカテゴリーには以下のようなものがある。（　）内にカテゴリーの略称を示す。

・学年

小学校から高校までを 3 学年ごとに分けたものを学年とした。小学校前半（小_前）・小学校後半（小_後）・中学校（中）・高校（高）の4種がある。

・教科

「教科書コーパス」の教科種別に従い、国語（国）・数学（数）・理科（理）・社会（社）・外国語（外）・技術家庭（技）・芸術（芸）・保健体育（保）・情報（情）・生活（生）の10種がある。教科書コーパスの教科種別は、小学校・中学校・高校の各学習指導要領（平成10～11年文部省告示、平成15年一部改正）に定める教科に基づいて設定されている。「教科書コーパス」の教科と各学習指導要領の教科の対応関係を表2に示す。

表 2　教科の対応関係

教科書コーパス	学習指導要領		
	小学校	中学校	高等学校
国語	国語	国語	国語
数学	算数	数学	数学
理科	理科	理科	理科
社会	社会	社会	地理歴史
			公民
外国語		外国語	外国語
技術家庭	家庭	技術・家庭	家庭
芸術	音楽	音楽	芸術
	図画工作	美術	
保健体育	体育	保健体育	保健体育
情報			情報
生活	生活		

また、媒体に関するカテゴリーとして、調査対象テキスト(1)の中学校・高校教科書部分を「教科書（教)」、調査対象テキスト(2)を「書籍（書)」と設定する。

４．語彙表の構成

収録対象語、計 95,286 語について、見出し語の五十音順に配列する。
PDF ファイルは次の 6 ファイルに分割する。

学校社会対照_ア.pdf	…ア行の文字で始まる見出し語を収録
学校社会対照_カ.pdf	…カ・ガ行の文字で始まる見出し語を収録
学校社会対照_サ.pdf	…サ・ザ行の文字で始まる見出し語を収録
学校社会対照_タナ.pdf	…タ・ダ・ナ行の文字で始まる見出し語を収録
学校社会対照_ハ.pdf	…ハ・バ・パ行の文字で始まる見出し語を収録
学校社会対照_マヤラワ.pdf	…マ・ヤ・ラ・ワ行の文字で始まる見出し語を収録

PDF ファイルの各ページに振られたページ番号は、6 ファイルの通し番号である。
語彙表は以下の①～⑦の項目によって構成される。

① 語彙素読み【語彙素-語彙素細分類】（品詞・活用型）〔語種〕

UniDic の語彙素読み・語彙素・品詞・語種を示す。語彙素は【　】内に、品詞は（　）内に、語種は〔　〕内に示す。語彙素細分類がある場合は、語彙素の後に続けて「-」を付して示し、活用型がある場合は、品詞の後に続けて「・」を付して示す。

② 初出学年

当該見出し語が調査対象テキスト(1)で初めて出現した学年を略称で示す。調査対象テキスト(1)で 1 度も出現しない場合は「-」で示す。

③ 度数　全・国・数・理・社・外・技・芸・保・情

「全」は調査対象テキスト(1)の中学校・高校教科書部分全体での度数を示す。

「国・数・理・社・外・技・芸・保・情」は調査対象テキスト(1)の中学校・高校教科書部分を 3.にあげた教科分類に従い 9 教科に分けた場合の、各教科での度数を示す。

度数が 0 の場合は「-」で示す。

④ レベル　LB・PB・PM・PN・OC・OY

調査対象テキスト(2)・(3)・(4)・(5)・(6)・(7)それぞれでのレベル[3]を示す。各テキストでの度数が 0 でレベルの設定できない場合は「-」で示す。

⑤ 特徴媒体

当該見出し語が「教科書」「書籍」のいずれで特徴語となるかをその略称で示す。調査対象資料(1)の中学校・高校教科書部分を当該資料、調査対象資料(2)を参照資料とした場合の特徴度[4]が 6.63（p>.01）より大きい場合、「教科書」の特徴語とみなし、-6.63 未満の場合、「書籍」の特徴語とみなす。どちらの特徴語でもない場合は「-」で示す。

⑥ 特徴教科

当該見出し語が特徴語となる教科の略称を示す。調査対象資料(1)の中学校・高校教科書の各教科部分を当該資料、調査対象資料(2)を参照資料とした場合の特徴度が 6.63（p>.01）より大きい場合、当該教科の特徴語とみなす。どの教科の特徴語でもない場合は「-」で示す。

⑦ 分類語彙表番号_分類語彙表見出し

当該見出し語と同語と認められる「分類語彙表 増補改訂版」データベース[5]のレコードの分類語彙表番号[6]と見出しを_で連結して示す。複数のレコードが同語と認められる場合、レコード間を;で連結する。

５．延べ語数・異なり語数

本語彙表に収録された見出し語 95,286 語のうち、調査対象テキスト(1)の中学校・高校教科書部分に出現するものは 48,077 語である。それらの語について、調査対象テキスト(1)の中学校・高校教科書部分での延べ語数・異なり語数を教科別に示すと表 3 のようになる。

---

[3] 付記 1 参照
[4] 付記 2 参照
[5] http://www.ninjal.ac.jp/products-k/kanko/goihyo/
[6] 「分類語彙表 増補改訂版」データベースのレコードを構成する項目のうち、分類番号・段落番号・小段落番号・語番号を-で連結したもの。

表 3　延べ語数・異なり語数

	中学校＋高校	
	延べ語数	異なり語数
全教科	2,232,037	48,077
国語	295,937	21,825
数学	174,618	4,209
理科	580,116	14,097
社会	697,321	26,192
外国語	51,374	5,693
技術家庭	158,744	10,273
芸術	134,836	14,070
保健体育	67,282	5,800
情報	71,809	4,342
生活	0	0

付記 1 レベルについて

レベルとは、調査対象テキスト(2)～(7)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（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	累積使用率（カバー率）
a	0　～　78%
b	～　88%
c	～　94%
d	～　97%
e	～　100%

付記 2 特徴度について

特徴度とは、当該資料における当該語が、他の資料（参照資料）と比べて出現度数の点でどの程度特徴的であるかを示す値である。特徴度には対数尤度比を補正した数値を用い、以下の式によって算出した（Kilgarriff、2001；内山・中條・山本・井佐原、2004）。

$$2(a\ln a+b\ln b+c\ln c+d\ln d-(a+b)\ln(a+b)-(a+c)\ln(a+c)-(b+d)\ln(b+d)-(c+d)\ln(c+d)+(a+b+c+d)\ln(a+b+c+d))$$[7]

a：当該資料での当該語の度数
b：参照資料での当該語の度数
c：当該資料の延べ語数－a
d：参照資料の延べ語数－b

※ただし、ad-bc<0 の場合、-1 を乗じる補正を行う。

特徴度が 0 であれば、当該資料と参照資料で当該語の出現の程度は等しい。特徴度が正の値で、かつ値が高ければ高いほど、当該資料において高頻度という意味で特徴的な語と見なされる。逆に、特徴度が負の値で、かつ値が低ければ低いほど、当該資料において低頻度という意味で特徴的な語と見なされる。

[7] ln は自然対数を表す。a または b が 0 の場合、alna または blnb を 0 として値を算出した（高見、2003）。

また、正の値の特徴度の有意水準とその臨界値は以下の通りである（高見、2003）。

有意水準	0.1（10%）	0.05（5%）	0.01（1%）	0.005（0.5%）	0.001（0.1%）
臨界値	2.71	3.84	6.63	7.88	10.83

文献

内山将夫、中條清美、山本英子、井佐原均（2004）「英語教育のための分野特徴単語の選定尺度の比較」 自然言語処理, 11-3, pp.165-197.

高見敏子（2003）「「高級紙語」と「大衆紙語」の corpus-driven な特定法」 （北海道大学）大学院国際広報メディア研究科・言語文化部紀要, 44, pp.73-105.

田中牧郎（2011）「語彙レベルに基づく社会的な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.

田中牧郎、近藤明日子、平山允子（2011）「教科書コーパス」 本報告書第 1 章第 1 節.

Adam Kilgarriff（2001）"Comparing corpora" *International Journal of Corpus Linguistics*, 6-1, pp.1-37.