日本語の学術共通語彙（アカデミック・ワード）の抽出と妥当性の検証

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

1．はじめに

本研究では，日本の大学の留学生など，日本語で学術テキストに触れる機会の多い学習者の学習効率を高めるため，日本語の学術共通語彙を抽出し，その妥当性を検討した。

学術共通語彙とは，一般的テキストでの使用率に比べて，学術的なテキストでより高い使用率を占める語彙であり，英語教育では幅広く知られている（Coxhead (2000)のAcademic Word List，Xue & Nation (1984) の University Word List など）。専門語彙が特定の分野においてのみ高い使用率を示すのに対し，学術共通語彙は分野を問わずに高い使用率を示す点が専門語彙とは異なる。いわば一般的な基本語彙と専門語彙の中間に位置する語彙であり，大学留学生にとっては，初級の基本的語彙に次いで重要な語彙である。

2．研究目的

第二言語・外国語の学習において，語彙学習の負担は非常に大きい。特に中級以降においては頻出する語句が減り，最頻出1000語では60～70%に達するテキストカバー率（以下，カバー率）も，それ以降では1000語ごとに数パーセントずつしかカバー率が上がらない（国立国語研究所1962など）。目的に即した効率的な語彙学習によって学習負担を軽減することは外国語・第二言語学習の重要な課題である。

留学生など，学術目的の学習者の場合，学術テキストに多く用いられる語彙に習熟する必要がある。目的が極めて限られている学習者の場合，初めから専門語彙を学ぶほうが効率的だと思われる(Ward, 1999)が，進学準備中など，専門を絞る前の段階では学術共通語彙を効率的に学ぶことが有効であろう。Tajinoほか(2010)は，カリキュラムの段階に応じた語彙リストのあり方を提案し，例えば，学術共通語彙，文系共通語彙を学び，そして経済学用語を学ぶというような階層を想定している。本研究もその考え方を踏襲している。

これまでにも初級語彙や専門語彙を含む留学生対象の語彙リストは多くあったが，学術語彙に絞られたものではなく，方法も主観を交えたものがほとんどであった。角(2010)やバトラー後藤(2010)など学術語彙的な性格を有する語彙リストもあるが，方法も対象も異なっており，方法や妥当性の検証において計量的裏付けがない。

ある学習・教育用の語彙リストの評価，特に効率性についての評価は，単位語数あたりのカバー率（延べ語数の使用率の合計）の高さによって示すことができる。リストの語が，対象テキストにおいて，リスト外の語より高い割合で出現するかどうかである。

本研究は，留学生などの語彙学習負担の軽減，より有効な語彙学習カリキュラムの開発のため，幅広い分野の学術的テキストにおいて一般的テキストより高いカバー率を示す語彙リストを作成し，カバー率の検証によって妥当性，有用性を検証することを目的とする。

3．研究方法

まず，『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年版）（国立国語研究所2009）の書籍部分約2800万語のテキストを，日本十進分類法などを手

がかりに人文系，社会系，理工系，生物・医学系の4領域に分類した。そのうえで，それぞれの領域についてCコード（出版社が発行時に付す販売対象コード）3000番台のテキストを専門テキスト（約300万語）として分別し，その他を一般テキストとした。

次に対数尤度比（log-likelihood ratio）(Dunning, 1993)を用いて，全領域の一般テキストにBCCWJのインターネットQ&Aサイト「Yahoo知恵袋」部分約500万語を加えたもの（計約3000万語）を参照テキストにして，各領域の専門テキストの特徴語を抽出した。対数尤度比は，特定の分布を要求せず，標本サイズが異なる場合にも適切な値を返し(Leech, Rayson, & Wilson, 2001, p 16)，適度な割合で特徴語を抽出する指標とされている(Chujo & Utiyama, 2006)，ので，本研究に適していると判断した。文理両面において使用される語を抽出するため，対数尤度比が3領域以上で正の値になる語をすべて抽出した。それを語彙データベース(松下 2011)の留学生用語彙ランク[1]によって表1のように分類した。結果として学術共通語彙として適切だとは思われないものも少数混ざっているが，恣意的に取り除くことはせず，今後，適切な基準によって改善しようと考えている。

カバー率の検証にはAntWordProfiler (Anthony, 2009)を用いた[2]。学術共通語彙の抽出の際に使用していないコーパスをテストコーパスとして，専門テキストにおいて，文芸作品，新聞，会話などの他のタイプのテキストより高いカバー率を示すかどうかを検証した。

テストコーパス　*冒頭の記号は表2の記号に対応する

(MC) 会話[3]：名大会話コーパス，約113万語（http://dbms.ninjal.ac.jp/nknet/ndata/nuc/ 2010年12月10日）

(BS) 一般書（文芸等[4]）：『現代日本語書き言葉均衡コーパス』2009年モニター版（国立国語研究所2009）「ベストセラー」部分，約230万語

(PC) 一般書（文芸等[5]）：「日英対訳文対応付けデータ」(内山・高橋 2003)の日本語部分，210万語（http://mastarpj.nict.go.jp/~mutiyama/align/index.html 2010年11月16日）

(JN) 新聞：日英新聞記事対応付けデータ（JENAAD）(Utiyama & Isahara, 2003)の日本語部分（1989-2001の「読売新聞」記事）約568万語

(IS) 人文・社会系専門テキスト：新屋映子・松下達彦編（未公刊）『日本語上級読解演習 国際学アラカルト』本文部分，約4万語

(TB) 社会系専門テキスト：「中・上級社会科学系読解教材テキストバンク」（東京外国語大学留学生日本語教育センター1998）本文部分，約19万語

(SS) 社会系専門テキスト：『留学生のための専門講義の日本語』（名古屋大学 国際化拠点整備事業2010）全9冊中，社会系の3冊分の講義テキスト部分，約5万語

(TN) 理工系専門テキスト：『留学生のための専門講義の日本語』（同上）全9冊中，理工系の5冊分

---

1 ランク付けの方法については松下(2010)参照。

2 形態素解析器にはMeCab（工藤 2006），解析用辞書にはUniDic（伝ほか 2009）を使用した。UniDicの出力をAntWordProfilerで使用するには加工が必要で，テキストエディタ上でマクロを作成して加工した。

3 科学研究費基盤 研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（平成13年度～15年度 研究代表者 大曽美恵子）の一環として作成された，120件合計約100時間の日本語母語話者同士の雑談を文字化した会話データ。本稿執筆時点ではダウンロードサイト不明。

4 テキスト数で文芸が53%を占め，他は各分野に分散しているが，専門テキストに分類されるものはない。

5 “Project Gutenberg” 「青空文庫」「プロジェクト杉田玄白」などの作品について，日本語文と英語文との対訳文対応を付けたもので，文芸書やエッセイの占める割合が高い。

の講義テキスト部分，約7万語

(BM) 生物・医学系専門テキスト：『留学生のための専門講義の日本語』（同上）全9冊中，生物・医学系の1冊分の講義テキスト部分，約1万語

表1　日本語学術共通語彙(JAWL)のレベル別異なり語数・語例・語種比率

学術共通語彙ラベル	旧日本語能力試験出題範囲	留学生のための一般的な語彙重要度ランク(*1)	レベル	4大領域中の共通領域数(*2)	異なり語数	語例（各カテゴリー中、一般的語彙重要度ランク最下位6語）	語種比率(%)（異なり語数）					
							和語	漢語	外来語	混種語	固有名詞	記号不明その他
JAWL 0	3級	679-1288	初級	4	31	科学　規則　割合 生産　産業　講義	25.8	67.7	0.0	3.2	0.0	3.2
				3	39	人口　スクリーン　数学 競争　工業　地理	20.5	71.8	7.7	0.0	0.0	0.0
JAWL I	2級 1級 級外	1289-5000	中級	4	559	発足　半数　配分 縮小　適正　見直し	20.6	74.6	2.1	2.5	0.0	0.2
JAWL II				3	542	演説　大小　実情 ステージ　ライフ　担保	14.2	76.8	6.5	1.3	1.1	0.2
JAWL III		5001-10000	上級前半	4	212	難問　能動　付随 定型　除　本稿	12.7	76.9	7.5	2.8	0.0	0.0
JAWL IV				3	452	交錯　カウント　精度 一因　箇年　エンド	12.4	75.9	9.1	1.5	0.9	0.2
JAWL V		10001-15000	上級後半	4	103	併存　親和　盛況 散在　補填　関わり合う	8.7	82.5	7.8	1.0	0.0	0.0
JAWL VI				3	328	帰着　編著　沿海 拮抗　常套　内情	13.1	75.0	9.5	1.5	0.3	0.6
JAWL VII		15000-20000	超上級	4	56	閉　増刊　含意 複　活路　所与	16.1	66.1	10.7	3.6	0.0	3.6
JAWL VIII				3	269	付則　深度　孤 概算　頒布　円錐	14.1	71.4	11.2	1.9	0.0	1.5

*1 松下(2010)による。旧日本語能力試験の4級と3級の語彙を優先的に上位に配置し，その他は書き言葉の頻度と分散度を掛け合わせた指標によりランク付けしている。

*2 人文，社会，理工，生物・医学の4大領域のうち，いくつの分野で特徴的かを示す。具体的には，専門テキストにおける対数尤度比（一般テキストを参照コーパスにした場合）が，上記4領域のうち，いくつの領域で＋となっているかを表す。

4．分析結果および考察

学術共通語彙リスト(JAWL = Japanese Academic Word List) 0からVIIIまでの9レベル，計2591語を抽出した（表1）。初級には学術共通語彙の語彙数も少なく，学習や教育を考える上では，中級のJAWL Iが最も重要なリストである。JAWL Iの559語は英語教育でよく知られているAcademic Word List (Coxhead, 2000)（以下AWL）570語に近い語数であるが，カバー率もAWLに非常に近い。抽出時に使用した学術コーパスでAWLは10.0%のカバー率を示したが，JAWL Iは11.1%である。

抽出時に使用していないテストコーパスでは，AWLが8.5%のカバー率であったのに対し，JAWL Iでは9.7～11.1%を示しており，一貫して高いカバー率を示している（表2）。

表２　日本語学術共通語彙のテキストカバー率・テキストカバー効率の比較

ジャンル				MC 会話 N=1129538		BS 一般書 N=2102178		PC 一般書 N=2298828		BCCWJ全体 N=32819424		JN 新聞 N=5675357		IS 人文・社会系 N=42152		TB 社会系 N=186768		SS 社会系 N=50601		TN 理工系 N=74645		BM 生物・医学系 N=13904	
語彙ラベル	旧日本語能力試験出題範囲	レベル	4大領域中の共通領域数（*1）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）	テキストカバー率（%）（*2）	テキストカバー効率（*3）
Basic (non-JAWL)	4級 3級			80.6	*660*	73.0	*602*	72.4	*586*	68.6	*552*	57.0	*466*	62.0	*991*	62.5	*722*	66.2	*1237*	59.8	*1154*	60.4	*2033*
JAWL 0	3級	初級	4	0.58	*187*	1.18	*382*	1.26	*405*	1.63	*525*	2.07	*667*	2.40	*799*	2.75	*888*	2.99	*998*	3.65	*1217*	2.73	*1139*
			3	0.65	*166*	1.39	*356*	1.31	*337*	1.38	*354*	1.85	*475*	2.38	*701*	2.35	*654*	2.22	*673*	1.80	*561*	1.65	*1029*
JAWL I	2級 1級 級外	中級	**4**	**0.77**	*17*	**3.14**	*56*	**2.66**	*48*	**4.57**	*82*	**8.70**	*156*	**10.2**	*221*	**9.72**	*178*	**9.77**	*237*	**11.1**	*279*	**11.1**	*457*
JAWL II			3	0.53	*12*	1.46	*27*	1.56	*29*	2.62	*48*	6.58	*122*	4.66	*126*	5.05	*99*	4.82	*161*	2.89	*115*	4.23	*365*
JAWL III		上級前半	4	0.03	*3*	0.12	*6*	0.07	*4*	0.16	*8*	0.32	*15*	0.23	*39*	0.37	*23*	0.38	*65*	0.68	*98*	0.45	*178*
JAWL IV			3	0.07	*3*	0.21	*5*	0.19	*4*	0.35	*8*	0.77	*17*	0.57	*42*	0.70	*24*	0.45	*58*	1.88	*169*	2.14	*521*
JAWL V		上級後半	4	0.01	*2*	0.02	*3*	0.02	*2*	0.03	*3*	0.05	*5*	0.06	*41*	0.05	*13*	0.05	*33*	0.08	*61*	0.06	*115*
JAWL VI			3	0.02	*2*	0.06	*3*	0.06	*2*	0.11	*3*	0.25	*8*	0.16	*35*	0.21	*15*	0.12	*55*	0.63	*126*	0.31	*221*
JAWL VII		超上級	4	0.00	*3*	0.00	*1*	0.01	*1*	0.01	*2*	0.01	*3*	0.04	*37*	0.02	*15*	0.02	*49*	0.10	*100*	0.04	*432*
JAWL VIII			3	0.01	*2*	0.03	*2*	0.03	*2*	0.05	*2*	0.11	*4*	0.05	*32*	0.08	*11*	0.07	*37*	0.38	*142*	0.05	*126*

*1 人文，社会，理工，生物・医学の４大領域のうち，いくつの分野で特徴的かを示す。具体的には，専門テキストにおける対数尤度比（一般テキストを参照コーパスにした場合）が，上記４分野のうち，いくつの分野で＋となっているかを表す。

*2 そのカテゴリーの語彙がテキスト全体に占める延べ語数の割合。

*3 テキストカバー率を，そのカテゴリに一の異なり語数で割り，1000000をかけたもの。そのカテゴリーの１種類の語が平均的にどれだけのテキストをカバーするかを示す。延べ語数と異なり語数の関係はテキストのサイズによって異なるため，同じテキストの中でカテゴリー間を比較することはできるが，大きさの著しく異なるテキスト間で数値を比較する際は注意が必要である。この数値が大きいほど，そのカテゴリーの語彙を学ぶことで効率よくそのテキストを理解できるようになることを予測する。

一方，非学術テキストでは，文芸書などの一般書でのカバー率 3%前後で，会話では 1%未満である。AWL では創作テキストで 1.4%のカバー率と報告されているが，本研究のテストコーパスがエッセイなども一部に含んでいることを考えれば近い数値と考えられるであろう。AWL は英語のリストであり，抽出時に基本語彙集である General Service List (GSL)(West, 1953)の 2000 語を除外しており，語の解析単位も JAWL I とは異なるため，単純な比較はできないが，初級語彙の占める割合との関係や，本研究のために案出したテキストカバー効率（そのカテゴリーの語を１語学習することで平均的にどのぐらいカバー率を上げられるかを示した指標，表２の注３を参照）の数値を見れば，JAWL は抽出方法においても妥当で，学術語彙の効率的な学習に適した語彙リストだといえる。

カバー率自体は上級以降極めて小さくなってなるが，JAWL の語彙は，他の語彙よりも効率よく学術テキストのカバー率を上げられる。上級以降でカバー率を 1%上げるために数千語も必要であることを考えれば，たとえ 0.1%でも効率よく学べることは重要である。

また，表２からは，新聞に初級語彙が少なく中級の学術共通語彙が多いことや，理工系や生物・医学系では人文系や社会系に比べて上級学術共通語彙が多いことがわかる。

JAWL II など，３領域語に欠けている１領域（一般テキストと比較して学術テキストで高い割合を示さなかった領域）を調べてみると，生物・医学系が 1630 語中 613 語（37.6%）と最も多く，以下，人文系 440 語（27.0%），理工系 343 語（21.0%），社会系 234 語（14.4%）となっている。このことは社会系学術テキストにおいて理工系や人文系との共通性が比較的高く，生物・医学系テキストは他の領域との共通性が相対的に低いことを示している。

学術共通語彙は意味的には抽象性が高い語が圧倒的に多い。「占める」「特殊」などの範

囲，「優れる」「属する」などの関係，「当初」「現状」などの段階，「減少」「強化」などの量的変化など論理操作に必要な語彙や，「取り上げる」「まとめる」など論述の展開に必要とされる語彙が非常に多い。3領域語には「署名」「保健」など具体的なイメージを伴う語が散見されるが，4領域語にはそのような語が少ない。語彙のレベルが変わっても，その性格は変わらないようである。意味・用法の面からも妥当なリストだと考えられる。

学術共通語彙の品詞を見ると，最も多いのは「形式」「背景」などの普通名詞で1072語（41.4%），次いで動名詞（＝サ変動詞語幹，スル動詞）が882語（34.0%）となっており，その他のタイプの名詞とあわせて2104語（81.2%）は名詞になり得る語である。動詞（動名詞を除く）が225語（8.7%），イ形容詞は「著しい」「等しい」など9語（0.3%）しかなく，「詳細」「平等」など，名詞にもなるナ形容詞（解析用辞書 UniDic では「形状詞」の一種）のほうが95語（3.7%）と多い。「-期」「-種」「各-」などの接辞が106語（4.1%）もあり，重要な位置を占めている。「しばしば」「あたかも」などの副詞は34語（1.3%）で，そのほかは助詞，助動詞，連体詞など22語(0.8%)のみである。これらの中には「のみ」「つつ」「べし」「あらゆる」「いかなる」「我が」「漠然」といった古語的な色彩を帯びた語が目立つ。受身・可能などを示す「れる」「られる」が選ばれていることも興味深い。

表1に戻って，語種比率（異なり語数）をみると，学術共通語彙では漢語が一貫して4分の3前後の高い割合を占める（全体では75.2%）。混種語も48語中「漢字一字＋する」の組み合わせ（「達する」「応ずる」「接する」など）が36語を占め，「概して」「総じて」「単に」などの副詞も漢語的性格が強いことを考え合わせると，学術共通語彙の77%程度が漢語系である。学術共通語彙の多くが明治期に創出されたいわゆる新漢語で，現代中国語との間で意味・用法のずれも小さいことを考えると，少なくとも学術テキストの語彙理解の点で，中国語系学習者には相当に有利な面があるといえる。和語は JAWL 0 と I で20%を超えるものの，それ以外では9〜16%程度にとどまっている。これは一般に和語の占める異なり語数の割合が高頻度の2000語レベルを除けばほぼ3分の1前後である（松下 2009, 2010）ことを考えると，かなり低い割合である。この語種比率の違いは大学などの学術的な場における日本語教育において，中国語系の学習者と非中国語系の学習者の語彙学習負担の違いに直結しており，カリキュラムを考える上でも重要な問題である。

5．今後の課題など

リストに含めるのに若干，不適切に見える語も少数だがある。対数尤度比が0より大きいという最低レベルを採録基準に採用しており，統計学的に有意でないであろうレベルまで落としているためだと思われる（Leech, Rayson, & Wilson (2001)は3.8を5%有意の基準としている）。主観によらず，何らかの基準で統一的に除去することを検討したい。

レベルづけにも恣意的な面がある。特に3領域しか条件を満たしていない語彙については，残りの一つを専門としている学習者にとっては一般的な重要性しか持たないことになる。例えばJAWLⅡの最下位より，JAWLⅢの最上位のほうが重要かもしれない。

本稿では触れられないが，作業の過程では2領域語，1領域語も抽出された。これを発展的にとらえると，学術テキストにおける語彙階層の全体像を明らかにできるものと思われる。また，本研究は実用的な観点から出発しているが，語彙的な側面から見たテキスト・

ジャンルあるいはレジスター変種の特徴づけにも利用できる。日本留学試験や大学の留学生入学試験など学術的であるべきテキストでは，学術共通語彙が一定の程度以上に含まれるようにテキストを選び，語彙をコントロールすべきであると考える。そして，当然のことながら，これらの語彙リストを如何に学習，教育に活用できるかを考えねばならない。

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

内山将夫・高橋真弓 (2003). 日英対訳文対応付けデータ.
http://www2.nict.go.jp/x/x161/members/mutiyama/align/index.html

小木曽智信 (2007). 「茶まめ」(形態素解析ユーザーインターフェース)
https://www.tokuteicorpus.jp/dist/modules/system/modules/menu/main.php?page_id=1&op=change_page

工藤　拓 (2006). MeCab Ver. 0.98pre3 (形態素解析器)　http://mecab.sourceforge.net/

国立国語研究所 (1962). 『現代雑誌九十種の用語用字 第一分冊 総記および語彙表』秀英出版

国立国語研究所 (2009). 「現代日本語書き言葉均衡コーパス」モニター公開データ2009 年度版

角　知行 (2010).「学術基本用語集作成の試み」『アカデミック・ジャパニーズ・ジャーナル』2, 11-21.

伝　康晴・山田　篤・小椋秀樹・小磯花絵・小木曽智信 (2009). UniDic version 1.3.11 (解析辞書)
http://www.tokuteicorpus.jp/dist/ (Version 1.3.0. first published in 2007)

バトラー後藤裕子 (2010). 「小中学生のための日本語学習語リスト (試案)」『母語・継承語・バイリンガル教育研究』6, 42-58.

松下達彦 (2009).「マクロに見た常用漢字語の日中対照研究　―データベース開発の過程から―」『桜美林言語教育論叢』5, 117-131.

松下達彦 (2010).「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―」『2010年度 日本語教育学会春季大会 予稿集』日本語教育学会, 335-336.

松下達彦 (2011). 日本語を読むための語彙データベース (Vocabulary database for reading Japanese) (= 日本語を読むためのTM語彙リスト Ver. 4.0). http://www.wa.commufa.jp/~tatsum/index.html

Anthony, L. (2007). AntConc Version 3.2.1 (text analysis tool)
http://www.antlab.sci.waseda.ac.jp/software.html (Version 1.0 first published in 2002)

Anthony, L. (2009). AntWordProfiler Version 1.2 w (word profiler)
http://www.antlab.sci.waseda.ac.jp/software.html (Version 1.0 first published in 2008)

Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System*, 34, 255-269.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19, 61–74.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Harlow: Longman.

Tajino, A., Dalsky, D., & Sasao, Y. (2010). Academic vocabulary reconsidered: An EAP curriculum-design perspective. *Iranian Journal of Teaching English as a Foreign Language and Literature*, 1(4), 3-21.

Utiyama, M. and Isahara, H. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. *ACL-2003*, 72-79.

Ward, J. (1999). How large a vocabulary do EAP Engineering students need? *Reading in a Foreign Language*, 12(2), 309-323.

West, M. (1953). *A General Service List of English Words*. London: Longman, Green & Co.

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication*, 3(2), 215-229.