JACET8000のテキストカバー率
語彙表と英文テキストによる検証*

上　村　俊　彦

The JACET List of 8000 Basic Words and Its Text Coverage: An Examination of Eight Word Lists and Four English Texts

Toshihiko UEMURA

Abstract: After gathering its own English texts, then comparing them with the British National Corpus, the JACET Committee for Revising Basic Words compiled a list of the 8000 most frequent and important English words. The list is available in a handbook *The JACET 8000 Basic Words*. In this article, the author, a member of this committee, examines to what extent words in this list correspond to those in four well-known academic word lists, and then tests how well these 8000 words actually cover the vocabulary in seven types of randomly chosen English texts.

1．はじめに

2003年３月，大学英語教育学会基本語改訂委員会による『大学英語教育学会基本語リスト：JACET8000』(1)が刊行された。（以下，JACET8000。ただし，JACET8000に掲載された語彙8000語自体に言及する時にはJ8Kと省略。）日本の英語学習者のために作成されたこの「基礎語彙表」は，今後，教育・研究の両面からさまざまな形で活用されることが期待される。

本稿では，さまざまな語彙表の掲載語がどの程度J8Kでカバーされるか検証するとともに，J8Kがさまざまな分野の英文テキストの語彙をどの程度カバーしているか検証する。

2．検証の概要

JACET8000(pp.122-124)には，J8K策定のために参照した各種の英英辞典定義語，アカデミック語彙リスト，既存語彙表などの語彙でJ8Kに採択されなかったものについての詳細な記述がある。また，望月(2003)は，英語学習者向けにもとの英文を制限語彙で書き直した英文テキスト（graded reader）を用いて，J8Kの英文テキストの「カバー率」の検証結果を報告している。なお，上村(2002)には，J8K（JACET8000刊行前の暫定版）を用いて，英語能力運用試験(STEP, TOEFL, TOEIC)問題に対する「カバー率」の検証に言及した部分がある。

本稿では視点を変えて，近年の語彙研究やコーパス研究でよく知られた語彙表や，インターネット経由で入手が容易にできる英文テキストを使ってJ8KとJ8KPlus（後述）の「カバー率」

＊ 本稿は，第42回(2003年度）大学英語教育学会全国大会JACET基礎語彙改訂委員会主催 シンポジュウム「JACET8000の活用と応用研究」の提案者として口頭発表したデータをもとに執筆した。

についての検証をおこなう。

2．1　語彙表

JACET8000付属のCD-ROMから，J8K単独，J8Kに別表（通称Plus250）を加えたJ8KPlus，以上2種類の語彙表を検証のために取り出す。検証データとなる語彙表は，CoxheadとWest（海外でよく知られた既存の語彙表），KLemma（British National Corpus（以下，BNC）のレマ・リスト），L1.2，L2.2，L2.3（BNCの頻度順語彙リスト3種類）と，「標準語彙水準12000」（日本で商用に開発されたALC社の語彙リストStandard Vocabulary List 12000。以下，SVL 12000。）とする。ただし，SVL 12000は，ALC8K（J8Kと語彙数をそろえたSVL Level 1～Level 8の8000語）とALC12K（SVL Level 1～Level 12の12000語）の2種類とする。

個々の語彙表（上記，初出時に太字で表示したもの）の詳細とTypeの数とを，表として**リスト1**にまとめて表示する。

リスト1．J8Kと検証用語彙表

略　称	語　彙　表	Type
ALC8K	ALC　SVL12000　Level 1～8	8036
ALC12K	ALC　SVL12000　Level 1～12	12036
Coxhead	A new academic word list	623
J8K	JACET 8000の本表	8049
J8KPlus	JACET 8000の本表 + Plus250	8254
KLemma	Kilgarriff lemma list	5512
L1.2	BNC rank frequency list for the whole corpus (not lemmatized) 最低出現頻度：100万語あたり75回	6723
L2.2	BNC rank frequency list: spoken English (not lemmatized) 最低出現頻度：100万語あたり10回	4295
L2.3	BNC rank frequency list: written English (not lemmatized) 最低出現頻度：100万語あたり20回	4303
West	A general service list of English words	2339

各語彙表の語数はType数で数えるため，品詞や語源の違いから語彙表では別の語彙として採録されていても，**リスト1**ではすべての同綴り語彙は1語として数える。また，**リスト1**の中で，ALC8K，ALC12K，J8K，J8KPlusの4つの語彙表は「米語綴り」，残りの語彙表は「英語綴り」のままとして，綴りの統一はおこなわなかった。

2．2　検証用英文テキスト

英文テキストの種類や分野などを考慮して，英文テキストによる検証作業は**リスト2**にある英文テキストデータを用いる。CNNテキストは，CNNのウェッブサイト約1日分。Horizonテキストは，BBCの科学番組Horizonの番組スクリプト12話分。Natureテキストは，8月14日付オンライン版記事から注記と文献リストを除いた論文本体。Reutersテキストは，Reuters Corpus CD-ROMの中の1ヶ月分（1997年3月1日～31日）。

リスト２．検証用英文テキスト

テキスト	ソース情報	Type/ Token
CNN.txt（74 files）	（書）（米）（時事　CNN オンラインニュース）	6,444/ 39,493
Horizon.txt（12 files）	（話）（英）（BBC TV 科学番組スクリプト）	6,643/ 72,746
Nature.txt（4 files）	（書）（? 英）（科学学術雑誌）	4,531/ 22,976
Reuters.txt（1951 Files）	（書）（英）（時事　CD-ROM）	22,169/ 492,806

なお，**リスト２**の（書）は「書き言葉」，（話）は「話し言葉」，（米）は「米語」，（英）は「英語」，（? 英）は版元（英），著者は多国籍であることを示す。

２．３　検証法

２．３．１　WordSmith とストップ・リスト

英文テキスト解析ソフトウェア WordSmith（以下，WS）を検証に使う。特定の語彙リストを「ストップ・リスト」として WS に登録し，検証用の既存語彙表あるいは英文テキストを WS にインプットして頻度順リスト出力コマンドを実行すると，そのストップ・リストに記載されなかった語彙のみの頻度順語彙リストと統計データ（Type や Token に関する数値データ）が得られる。

個々のストップ・リストは，使用する語彙表の Typeすべてに，(1)アルファベット大文字（A, B, ... Z），(2)BE, DO, HAVEの活用形（IS, AM, ARE, WAS, WERE; DOES, DID; HAS, HAD），(3)短縮形（'M, 'RE, 'D, 'VE, 'S 等）を加えたものから構成される。

なお，ストップ・リストとして使用する語彙表の「米語綴り」，「英語綴り」についてはそのままとして修正や統一をおこなっていない。

２．３．２　カバー率

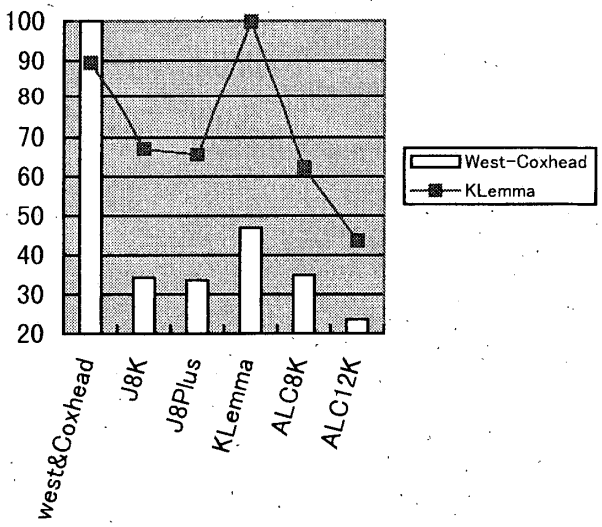
語彙表や英文テキストの「カバー率」計算は，WS のストップ・リスト未設定時（初期設定）の Type 総数と，ストップ・リスト設定時（設定変更後）の Type 数をもとに計算する。WS のストップ・リスト未設定時に出力された Type 総数をA，設定時に出力された Type の数をBとすると，「カバー率」の算定式は（A－B）÷A×100となる。すなわち，ストップ・リストに一致した語彙の Type 数を総 Type 数で割り，百分率を取ったものが「カバー率」である。

３．語彙表とカバー率

それぞれの語彙表が，ほかの語彙表の語彙をどのくらいカバーしているのか，語彙表相互のカバー率を調べた。**グラフ１**は，West-CoxheadとKLemmaとの他の語彙表に対する「カバー率」を示す。なお，「カバー率」100%表示からも明らかなように，**グラフ１**の棒グラフは West-Coxhead の，折れ線グラフは KLemma の「カバー率」である。ただし，アカデミック語彙表として知られる Coxhead は，West を「一般語彙」（general vocabulary）として習得していることを前提としているので，両者の掲載語（623語＋2339語）全体で１つのストップ・リストとした。

West-Coxheadの「カバー率」は，自己に対するものを除くと，最高は KLemma に対する47%で，それ以外は35%以下となった。KLemmaの「カバー率」は，最低が ALC12K に対する43.6%，最高は West & Coxhead に対する90%，その他は62.1%～67%となった。この結果をまとめた**グラフ１**からも明らかなように，West-Coxhead（Type 数2962語）に存在しない語彙が他

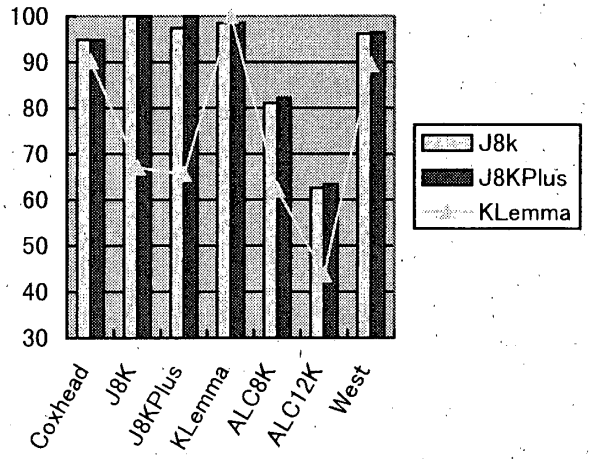
グラフ1．West-Coxhead と KLemma



の語彙表に多数採録されている。一方，KLemma（Type 数5512語）の場合，ALC12K（Type 数12000語）を除くと，語彙の62％以上は同一となり高い重なり率が認められる。ただし，KLemma が BNC，JACET8000のレマ・リストとして使われたこと，ALC の SVL12000の策定に BNC データが重要な判断基準(2)として利用されたことを考慮すると，60％を超える語彙の一致は当然の結果と判断できる。

以下の**グラフ 2** は J8K と J8Plus，**グラフ 3** は ALC8K と ALC12K の「カバー率」を示している。（なお，比較のために，KLemma の「カバー率」を折れ線で表示している。）

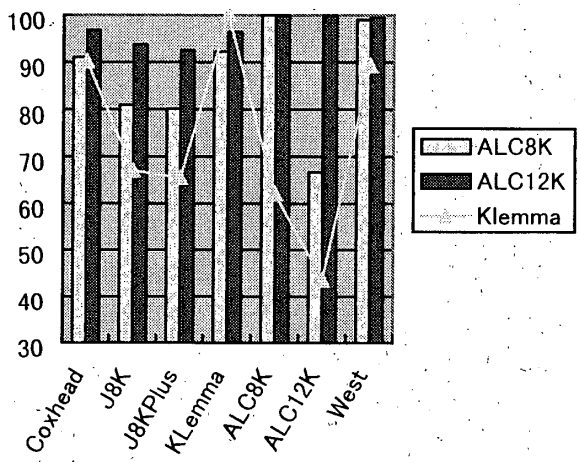
グラフ2．J 8 K ＆ J 8 KPlus



グラフ 2 と**グラフ 3** からも明らかなように，Coxhead，KLemma，West に対する J8K と J8kPlus の「カバー率」，あるいは ALC8K と ALC12K の「カバー率」はともに90％を超えている。しかし，**グラフ 2** では，J8K と J8KPlus の ALC8K に対する「カバー率」は80％台，ALC12K に対しては60％台のカバー率となっている。

同様な傾向は**グラフ 3** についても見られる。すなわち，ALC8K はほぼ同数の語彙数で構成される J8K，J8KPlus に対して80％台の「カバー率」を示している。以上，J8K と ALC8K とは相

グラフ3．ALC8K & ALC12K



互に80%台の「カバー率」であることから，2つの表の8000語中，約6400語程度は一致していると推測される。

なお，ALC8Kに4000語Typeの数が増えたALC12Kの場合，J8K，J8KPlus両者に対する「カバー率」は93%前後に上昇している。しかし，見方を変えると，ALC12K（=SVL12000）でも，J8KやJ8KPlusの語彙の約500語から600語はなおカバーされずに残っていることになる。

グラフ2，**グラフ3**によると，J8KとJ8KPlus，あるいはALC8KとALC12KのCoxheadとWestに対する「カバー率」は，J8K（Coxhead 94.9%，West 96.2%），J8KPlus（Coxhead 94.9%，West 96.5%），ALC8K（Coxhead 91.1%，West 99.0%），ALC12K（Coxhead 96.8%，West 99.4%）である。この高い「カバー率」から判断すると，J8KとALC8Kとは，ともにCoxheadとWestの語彙の大部分を含んでいるものと考えられる。このことが意味することについては，後に**5**で検討する。

Leech, Rayson & Wilson(2001:1)によると，BNCは約1億語相当の英文テキスト(3)からなり，その中に異なり語（different word form）は757,087語ある。ただし，BNCに10回以上出現する語形（word form）は124,002語である。（同掲書 p.9）

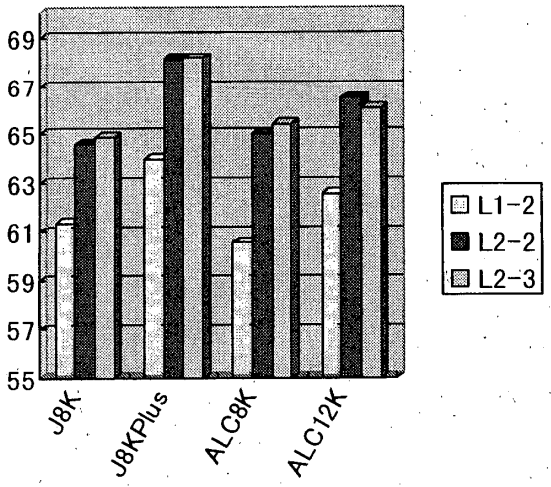
3つのBNC語彙表（L1-2，L2-2，L2-3）は，BNCの「全体の」頻度順リスト，「話し言葉」の頻度順リスト，「書き言葉」の頻度順リストから，特に出現回数の多い語彙を抽出して作成された語彙のリストである。

J8K，J8KPlus，ALC8K，ALC12Kの掲載語彙をストップ・リストとして，3つのBNC語彙表に対する「カバー率」を調べた結果をまとめ，**グラフ4**とした。（ただし，**グラフ4**ではストップ・リストをX軸とした。）

BNC語彙表（L1-2，L2-2，L2-3）に対する，8000語レベルのJ8K，J8KPlus，ALC8Kの「カバー率」と，12000語レベルのALC12Kの「カバー率」との間には，**グラフ4**からも明らかなように大きな相違は認められなかった。ただし，J8KとJ8KPlus，ALC8KとALC12Kとの「カバー率」比較では，語彙数の多いJ8KPlusとALC12Kが，8000語レベルのJ8KPlusと12000語レベルのALC12Kとの「カバー率」比較では，J8KPlusのほうが若干高めの数値となった。ALC12Kは，8000語レベルの語彙表に4000語を加えたものであるが，1億語レベルの英文テキストにもとづくBNC語彙表に対するカバー率が8000語レベルの語彙表とあまり違わない結果となったことは予想外であった。

J8KとBNC語彙表の語彙の一致を，J8Kの「カバー率」L1.2（61.34%），L2.2（64.5%），

グラフ４．L1-2，L2-2 & L2-3

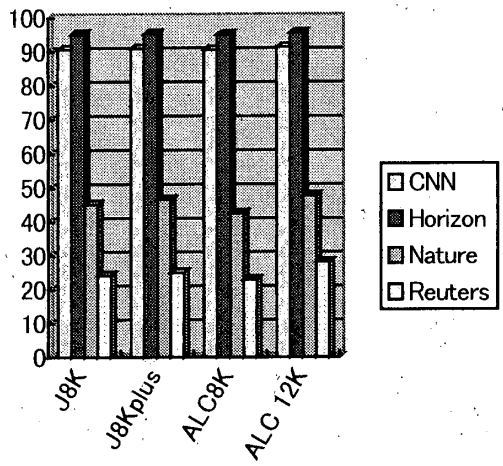


L2.3（64.88%）を参考に計算すると，J8K語彙と一致するのはL1.2で4124語，L2.2で2773語，L2.3で2792語となった。（語彙数については，小数点１位を四捨五入。）

４．英文テキストと「カバー率」

４つの英文テキスト（CNN，Horizon，Nature，Reuters）に対するJ8K，J8KPlus，ALC8K，ALC12Kの「カバー率」を出して，**グラフ５**とした。（X軸がストップ・リスト）

グラフ５．検証英文テキスト



４つのストップ・リストの「カバー率」の最小値と最大値を並べると，CNN（90.6%，91.3%），Horizon（95.1%，95.6%），Nature（42.5%，47.8%），Reuters（22.9%，28.1%）となり，語彙表の違いによる「カバー率」の相違は，CNNとHorizonとは１%前後，Natureは約５%，Reutersは約６%程度となった。8000語レベルの語彙表（J8K，J8KPlus，ALC8K）の「カバー率」最高値と，12000語レベルのALC12Kの「カバー率」とを比較すると，CNN（0.3%），Horizon（0.3%），Nature（1.2%），Reuters（3.1%）にとどまった。8000語レベルと12000語レベルの語彙表による「カバー率」の比較は，前節３における各語彙表相互の「カバー率」比較でも試みたが，語彙数4000語の積み上げが今回も「カバー率」の大きな改善に貢献していないことが明らかとなった。

4テキストのうち，CNNとHorizonに対する4つの語彙表の「カバー率」は90%以上と高い数字となったが，Natureで40%台，Reutersで20%台と低い「カバー率」にとどまった。科学や時事分野に多く見られる数字，地名・人名などの固有名詞，科学関連専門用語，頭辞語を含む時事英語特有の語彙などは，本来，J8KやALC8Kのような英語学習者向け語彙表のカバー範囲から外れている。低い「カバー率」にとどまったのは，NatureとReutersの英文テキストには，このような対象外の語彙が多く使われていることが大きな要因となったと思われる。(4)

なお，語彙表の「英語綴り」「米語綴り」については，英文テキストについても元データのままとし，いかなる編集もおこなっていない。両英文テキストの出版元は英国にあることを考慮すると，英文テキストの「英語綴り」も若干の影響となった可能性がある。（2．1参照。）

5．Nation(2001)の4000語と「カバー率」

Nation(2001:147)によると，英文テキストを楽に読みこなすためには，全テキストに占める既知の語彙比率は98%以上であることが好ましい。また，Nationによると，「教科書」(academic textbook）の英文テキストについて，その中の語彙が95%まで既知であるためには，4000語レベルの「ワードファミリー」(5)が必要である。なお，その内訳は，極めて出現頻度の高い一般語彙2000語（WestのGeneral Service Listに該当（Nation(2001:15-16)），一般教養語彙（CoxheadのAcademic Word List掲載の570語はこの範疇），専門語彙1000語以上，その他に固有名詞や出現頻度は低いが分野に特有な語を想定している。(6)

すでに**3**の**グラフ2**と**グラフ3**の考察で述べたように，J8KとALC8KはCoxheadとWestの語彙をほとんどカバーしている。このことは，Nationの意図した基礎語（2000語）と一般教養語（570語）は，そのほとんどがJ8KとALC8Kに含まれていることを意味する。ただし，留意すべき点は，J8KとALC8Kは語彙の頻度順リストであるのに対して，CoxheadとWest中の語彙は単なる項目リストであることから，両語彙表の個々の語彙は出現頻度順8000語リストの中に分散して存在していることである。換言すると，このような語彙約2600語は，8000語リストの頻度上位3000レベルまでにすべて入っているとは限らない。

なお，ストップ・リストでカバーされなかったCoxheadの「英語綴り」語彙（labour，maximise，minimise，utilise）は，「米語綴り」(labor，maximize，minimize）でJ8Kの中に含まれている。

6．おわりに

JACET8000から出力された2つの語彙表（J8KやJ8KPlus）のテキスト「カバー率」の検証を，さまざまな語彙表や4種類の英文テキストを用いておこなった。J8KとJ8Kに250語を加えたJ8KPlusは，(1)Nationの「基本語2000」と「一般教養語」に対しては95%前後，(2)ALC8Kに対しては80%程度，(3)BNCの出現頻度が特に高い語彙からなる3種類のリスト(L1-2，L2-2，L2-3)に対しては60%以上，(4) 4つの英文テキストの中で，専門性の低い一般的な英文テキストに対しては90%以上，という高い率でそれぞれの語彙をカバーしていることが明らかとなった。

J8Kや英語導入期の学習者に配慮したJ8KPlusは，多岐にわたる英文テキストを集めたサブ・コーパスの構築，その頻度順語彙リスト作成，主要な各種語彙表の語彙調査，BNC語彙の出現頻度順リストを用いたサブ・コーパス語彙の頻度順リストの頻度補正などの一連の作業をもとに作り出された。本稿の検証からも明らかなように，J8KやJ8KPlusは英語学習者のための語彙表として信頼に値するものとして評価できる。

ただし，一連のWSのストップ・リストを使った検証作業では，J8KやJ8KPlusの採録語彙

の出現頻度順位についての妥当性の検証はおこなっていない。J8KやJ8KPlusは，自前のコーパスによる語彙の出現頻度とBNCの出現頻度順語彙リストとの差分を取る補正作業を経て作り上げられている。頻度順位に関する正当な検証作業は，BNC以外の大規模でデータバランスの取れた英文テキスト・コーパスが利用できて初めて可能となる。今後の課題としたい。

注

(1) 日本人英語学習者にとって重要な英文テキストを網羅的に集め，品詞タグを付けた後に，出力された頻度順語彙リストがベースとなっている。J8K確定作業には，現在，もっとも大規模な英語テキストコーパスの1つであるBritish National Corpus（以下，BNC）でも使われた品詞タグセットCLAWSやKilgarriffのレマ・リスト（本稿では，KLemmaと略記）を利用するなど，BNC作業手順を参考にしておこなわれた。

(2) ALCのホームページ中の「SVLの成り立ちSTEP2」に，「Adam KilgarriffのBritish National Corpus frequency listを見出し語としたものをベースに，各種コーパス，学習英和辞書の重要語表示などを参考にしつつ，複数のアメリカ人のネイティブスピーカーの感覚的判断を記号化して振り分け，各単語のレベルを決定する。」という表記がある。（http://www.alc.co.jp/goi/PW_top_all.htm）本稿のKLemmaも上記のKilgarriffのリストから派生したもの。

(3) BNC以外の国際的な英文テキストコーパスの1つがICE(International Corpus of English)。このプロジェクトでは，英語圏の22の国または地域ごとに100万語レベルのコーパス（500テキスト。ただし，1テキストは2000語）を作る国際的なコンピュータ・コーパス作成プロジェクト。BNCとICE-GB（英国版ICEコーパス）のコーパス・データ量を比較すると，100対1となる。

(4) 以下は，Reutersテキスト（1997年3月31日付け記事の1つ）の英文テキスト部分を中心として抜粋したもの。（太字は人名，地名，数字等を示す。）

- <title>UK: SOCCER-ITALY FACE CRUCIAL TEST AGAINST POLAND.</title>
- <headline>SOCCER-ITALY FACE CRUCIAL TEST AGAINST POLAND.</headline>
- <byline>Barney Spender</byline>
- <dateline>LONDON 1997-03-31</dateline>
- <text>
- Christian Vieri had the honour of celebrating his international debut on Saturday by scoring Italy's 1,000th goal but another against Poland in Chorzow on Wednesday could have the greater long-term significance.
- Having already beaten England at Wembley, a win for Cesare Maldini's men would effectively secure their place at next year's World Cup finals in France.
- It would put them six points clear at the top of European group two with home games against their main challengers still to come.
- </text>

(5) Nation(2001:8)によると「ワードファミリー」とは，見出し語(headword)，その活用形やさまざまな派生形が含まれる。

(6) Hirsh and Nation(1992)suggest that for ease of reading where reading could be a pleasurable activity, 98-99% coverage is desirable (about one unknown word in every 50-100 running words). To reach 95% coverage of academic text, a vocabulary size of around 4,000 word

families would be needed, consisting of *2,000 high-frequency general service words, about 570 general academic words* (the Academic Word List) and *1,000 or more technical words, proper nouns* and *low-frequency words*. (Nation(2001:147)) *(下線上村)*

なお，上記英文の下線部 academic text は，"The text is from an academic textbooks..." (同掲書 P12) から「教科書」と解釈される。

参照文献

(和文)

大学英語教育学会　基本語改訂委員会（編）(2003)『大学英語教育学会基本語リスト JACET 8000』東京：大学英語教育学会

上村俊彦(2002)「3つの英語試験と7つのウェッブ英語テキストの語彙研究」県立長崎シーボルト大学　国際情報学部紀要　第3号　pp. 179-190

(英文)

Leech, G.; Rayson, P. & Wilson, A. (2001) Word *frequencies in written and spoken English based on the British National Corpus.* Harrow: Pearson Educational Ltd.

Mochizuki, M. (2003) "JACET8000 compared with other vocabulary lists" in Murata, Yamada & Tono eds. *AsiaLex '03 Tokyo proceedings.* pp. 378-383.

Nation, I.S.P. (2001) *Learning vocabulary in another language.* Cambridge: CUP.

Nelson, G.; Wallis, S. & Aarts, B. (2002) *Exploring natural language: Working with the British component of the International Corpus of English.* Amsterdam: John Benjamins Publishing Company

語彙表*

ALC　Standard vocabulary list 12000
http://www.alc.co.jp/goi/PW_top_all.htm

Coxhead, A. (2000). A new academic word list
http://www.vuw.ac.nz/lals/div1/awl/awlinfo.html

Kilgarriff, A. (1998) BNC database and word frequency lists. HTML version:http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

Leech, G., Rayson, P & Wilson, A. (2001) *Word frequencies in written and spoken English based on the British National Corpus.* Edinburgh: Pearson Educational Limited
http://www.comp.lancs.ac.uk/ucrel/bncfreq/

Reuters corpus. Vol.1 English language, 1996-08-20 to 1997-08-19
http://www.reuters.com/researchandstandards/corpus

West, 1953, *A general service list of English words,* London: Longman.

検証英文テキスト情報

(1) BBC Horizonテキスト

H120299.txt (The Midas Formula, BBC2 9:30pm Thursday 2nd December 1999)

---

* J8K と J8KPlus については，『大学英語教育学会基本語リスト　JACET8000』付属 CD-ROM から取り出したもの。

http://www.bbc.co.uk/science/horizon/1999/midas_script.shtml
H040400.txt（Moon Children, BBC2 9:00pm Tuesday 4th April 2000）
http://www.bbc.co.uk/science/horizon/1999/moonchild_script.shtml
H102600.txt（The Lost World of Lake Vostok, BBC2 9:00pm Thursday 26th October 2000）
http://www.bbc.co.uk/science/horizon/2000/vostok_transcript.shtml
H022201.txt（Snowball Earth, BBC2 9.00pm Thursday 22nd February 2001）
http://www.bbc.co.uk/science/horizon/2000/snowballearth_transcript.shtml
H03H0801.txt（Taming The Problem Child, BBC2 9.00pm Thursday 8th March 2001）
http://www.bbc.co.uk/science/horizon/2000/problemchild_transcript.shtml
H030701.txt（The Fall of the World Trade Center, BBC Two 9.00pm Thursday 7 March 2002）
http://www.bbc.co.uk/science/horizon/2001/worldtradecentertrans.shtml
H092001.txt（The Mystery of the Persian Mummy, First shown: BBC Two 9.00pm Thursday 20 September 2001
http://www.bbc.co.uk/science/horizon/2001/persianmummytrans.shtml
H022102.txt（The Dinosaur that Fooled the World, First shown: BBC Two 9.00pm Thursday 21 February 2002）
http://www.bbc.co.uk/science/horizon/2001/dinofooltrans.shtml
H031402.txt（Archimedes' Secret, BBC Two 9.00pm Thursday 14 March 2002）
http://www.bbc.co.uk/science/horizon/2001/archimedestrans.shtml
H051602.txt（The A6 Murder, BBC Two 9.00pm Thursday 16 May 2002）
http://www.bbc.co.uk/science/horizon/2001/a6murdertrans.shtml
H101202.txt（Mega-tsunami: Wave of Destruction, BBC2 9:30pm Thursday 12th October 2000
http://www.bbc.co.uk/science/horizon/2000/mega_tsunami_transcript.shtml
H050303.txt（Flight 587, BBC2 9:30pm Thursday 30th May 2003）
http://www.bbc.co.uk/science/horizon/2003/flight587trans.shtml
(2) CNN テキスト　オンラインニュース http://www.cnn.com（2003年5月15日）
(3) Nature テキスト　Nature 424（2003年8月14日付け）
(4) Reuters corpus. Vol. 1 English language, 1996-08-20 to 1997-08-19 CD-ROM（1997年3月1日～31日）