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

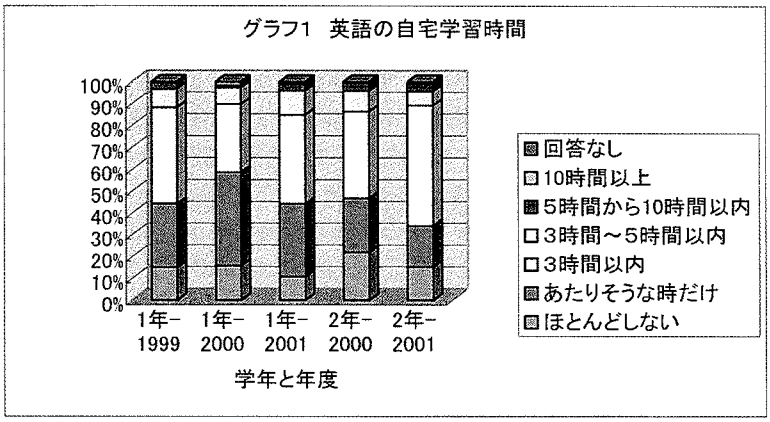
上　村　俊　彦

A Corpus Based Study of English Words in Three English Proficiency Tests and Seven Web English Texts

Toshihiko Uemura

1．はじめに

3年に渡る「シーボルト大学学生の外国語学習実態調査」（以下，「実態調査」）から，シーボルト大学（以下，本学）学生の8割は，自宅における1週間当たりの英語学習時間が3時間以下であることが明らかとなった。[1)]



グラフ1は，「実態調査」をもとにして，本学学生の自宅学習時間をまとめたもの。ただし，「実態調査」は，本学の開学年（1999年）は1年生のみが対象，2000年と2001年については1年生と2年生が対象となった。

自ら英語を学ぶことに消極的な学生に自発的な英語学習を促すために，本学では英語能力検定試験（実用英語技能検定（STEP試験），TOEIC試験，TOEFL試験）で一定以上の成績をあげた学生にはその成績を単位認定したり，本学の外国語学習支援ホームページに学生向けの英語学習リソースのリンク情報[2)]を掲載するなどの試みをおこなっている。

このような指導の成果か，3つの英語試験の共通点や相違点，ウェッブサイト上の新聞社や放送局などが発信する英語情報の利用法や英語テキストとしての難易度などについて，学生から質問を受けるようになってきた。

近年，コーパス言語学の研究は大いに進み，Hunston(2002)やMeyer(2002)からも明らかなように，さまざまな英語テキストをデータとした英語の語彙の研究方法が確立されてきた。本稿では，コーパス言語学の手法を用いて，学生がしばしば接する英語能力試験問題やウェッブサイト

上の英語の「語」について，その出現傾向や英語学習語彙としての重要性などについて検証する。

２．データ収集と処理

２．１　STEP, TOEIC と TOEFL

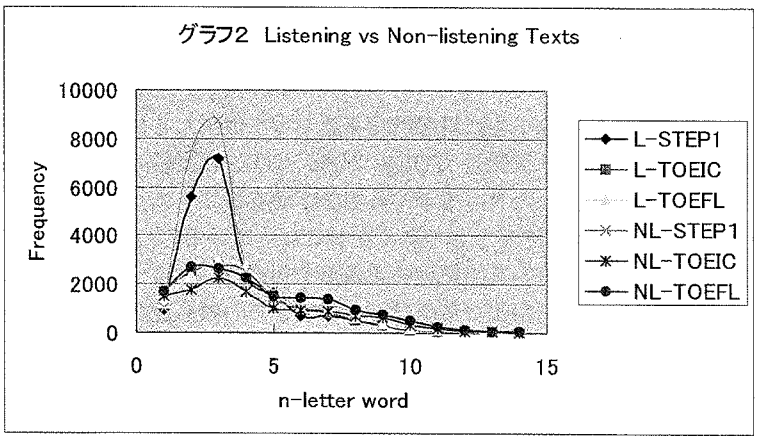
学生に広く認知されている英語試験の中から，英検１級(STEP1)，TOEIC，TOEFL の試験問題をコーパスデータとする研究を行う。３種類の試験問題の過去問題，あるいは（公認）想定問題の３回分をコンピュータによるデータ処理解析ができるように，スキャナーと英文テキスト変換ソフト OmniPage を使ってテキストファイル形式の電子ファイルとした。ただし，英検１級のスピーチ課題，TOEFL の Writing 問題は対象外とした。

試験問題データは，リスニング部とそれ以外の部分とにそれぞれ切り分た。データ処理は，「リスニング問題」３ファイル，「リスニング以外の問題」３ファイルのグループごとに Word-Smith という英文テキスト解析ソフトを使った。**Appendix 1**は，その出力結果を整理したものである。この表によると，出力された「語」の数（Types)，「語」を構成する文字数(Average Word)，「文」を構成する「語」数（(Standardized) Sentence Length）のすべてにおいて，「リスニング問題」よりも「リスニング以外の問題」の数値の方が大きくなった。通常，「話し言葉」よりも「書き言葉」に近いテキストのほうが，これらの数値は大きくなる。「リスニング以外の問題」は，明らかに「リスニング問題」よりも「書き言葉」に近いテキストであり，出力された数値はこの傾向を示している。

３つの試験問題テキストを比較すると，テキスト中の１語あたりの平均文字数が多い（すなわち，長いスペルの語）が最も多く見られたのは TOEFL であった。また，テキストに現れた「語」の種類が最も多く，文を構成する「語」の数が最も多い（すなわち，最も長い文の）テキストは STEP1 であった。換言すると，「語彙」の視点から３つの試験問題テキストを見ると，STEP1 テキストが英語学習者に最も「語彙」力を要求するテキストであったということになる。

Appendix 1のデータをもとに，テキストの「語」が何文字から構成されているか，14段階（１文字語から14文字以上の「語」）に分けてグラフ化したのが**グラフ２**である。

「語」の出現頻度は，３文字語前後でピークとなり，その後は右下がりに数値が減少する分布パターンとなっている。「リスニング問題」，「リスニング以外の問題」の各テキストファイルは，基本的にこの分布傾向を示している，ただし，STEP1 の２文字語と３文字語の数値は，他のテキストの数値を大きく引き離して高い山となっている。



（ただし，**グラフ２**の **L** は「リスニング問題」，**NL** は「リスニング以外の問題」の略記。）

WordSmith から出力された語の出現頻度リストを見ると，特徴のある傾向を示す動詞があった。出現頻度順位が「リスニング問題」の上位100位以内にありながら，「リスニング以外の問題」ではその順位から100位以上の降下が見られた動詞は，hear(63/2)　think(58/6)　look(56/11)　know(55/12) であった。(括弧内は，「リスニング問題」,「リスニング以外の問題」それぞれの出現回数。)

データ1　「リスニング」の出現回数

	STEP1	TOEIC	TOEFL	total
hear	1	24	38	63
think	17	26	16	58
look	4	43	9	56
know	19	8	28	55

なお，**データ1**は，「リスニング問題」3つのファイルにおける，テキストごとの出現回数を示す。

WordSmith のコンコーダンス機能によると，think または know と頻繁に共起する「語」が確認できる。think の用例(58例)中，think と共起する主語で出現頻度が高いのは，I(14例)，You（または you）(13例)，3人称単数形主語(8例)である。また，疑問文(13例)中，12例はdo you と think とが共起するパターンである。

think　I *think* it looks...(TOEIC) / Do you *think* they'll be...(TOEFL)

know の用例(55例)中，know と共起する主語で出現頻度が高いのは，I(15例)，You（または you）(13例)，3人称単数主語(16例)である。なお，3人称単数形主語＋doesn't（あるいは didn't）＋know という共起パターンは12例である。

know　I *know* where that one is...(TOEIC) / He doesn't *know* how to paint...(TOEFL)

また，hear はそのほとんどが問題指示文の中の表現で，テキスト中の一般的な用例は少なかった。同じく，look は TOEIC 問題の指示文の表現で，テキスト中の一般的な用例ではなかった。

hear　On the recording, you *hear*...(TOEFL)

look　*Look* at the picture mark...(TOEIC)

逆に，出現度順位が「リスニング以外の問題」の上位100位以内にありながら，「リスニング問題」ではその順位から100位以上の降下が見られた動詞は，used(18/48)　refer(10/44)　united(5/44)　choose(23/43)であった。(括弧内は，「リスニング問題」,「リスニング以外の問題」それぞれの出現回数。)

データ2　「リスニング以外」の出現回数

	STEP1	TOEIC	TOEFL	total
used	11	5	32	48
refer	2	41	1	44
united	3	4	37	44
choose	10	9	24	43

なお，**データ2**は，「リスニング以外の問題」3つのファイルにおける，テキストごとの出現回数を示す。

used はそのほとんどが動詞で，準助動詞（used to）の例は2例のみであった。

used　...engineers have *used* computer graphics...(TOEIC)
This school *used* to have...(STEP)

united は，本動詞としての用例１件を除き，*United* States（40例）*United* Nations（３例）のように複合語の第１要素としての用法であった。

また，以下の例が示すように，refer は TOEIC の，choose は３つの試験テキストの問題指示文や解答例の中の表現で，テキスト中の一般的な用例ではなかった。

refer　Questions 195-196 *refer* to the following...(TOEIC)

choose　Therefore, you should *choose* (D). Now begin...(TOEFL)

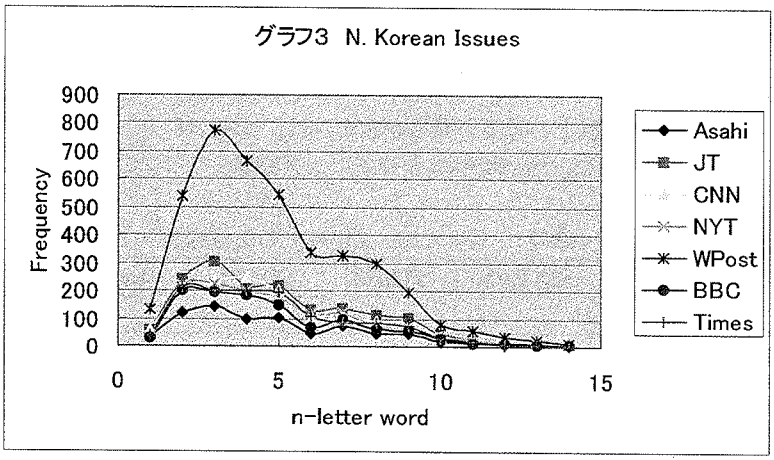
２．２　インターネットの英語情報

インターネットは，学習や研究に活用できる英語情報の宝庫である。しかも，通常の学術書や雑誌記事と比べて，インターネット情報には速報性がある。日頃から，学生には英語学習や情報収集の際の重要なソースとして，国内外の新聞社や放送機関のウェッブサイトの閲覧を勧めている。

今回，ウェッブ上の英文テキストの例として収集したのは，９月17日の日朝首脳会談後に報道された北朝鮮による日本人拉致事件に関する記事である。任意に選んだ日本，米国，英国のマスメディア関連ウェッブサイト７箇所に9月21日にアクセスし，それぞれの英文記事を集め，テキストファイル形式の電子ファイルとした。**Appendix 2** は，７つの電子ファイルを WordSmith に入力して得られた結果を整理したものである。

日本サイト（Asahi･JT），米国サイト（CNN･NYT･WPost），英国サイト（BBC･Times）の各英語記事テキストの中で，「語」を構成する文字数が最も大きかったのは NYT(5.05)，「文」あたりの語数が最も多かったのは Times(15.22)であった。

Appendix 2 のデータをもとに，テキストの「語」が何文字から構成されているか，14段階に分けてグラフ化したのが**グラフ 3** である。頻度のピークは３文字語近くで，その後は右下がりに数値が減少している。ただし，７つのテキストの中で，WPost の２文字語から９文字語までの出現頻度の数値は，データ量が多いために他の６テキストの数値よりも高めに位置している。



WordSmith の「語」出現頻度順リストに頻度20回以上で出現するものの多くは，どのような英文テキストの「語」出現頻度順リストでも上位を占める機能語が中心であった。ただし，７つの北朝鮮拉致問題テキストでは，関連の国名・地名・人名などの固有名詞：Korea(147) Japanese(145) Kim(110) Japan(94) Korean(83) Koizumi(73) Pyongyang (41) Tokyo(27)

South(22) States(22) Jong (21) United(20)や，会談内容の記述に関連する名詞・形容詞・動詞：talks(41) died(30) leader(27) relations(27) summit(27) abducted(26) alive(26) missing (25) families(24) abductions(23) kidnapped(23) dead(22) minister(22) normalization(22) news (21) people(21) missile(20) promised(20) などの「語」が高い頻度となって現れている。ちなみに，出現回数20回以上の「語」は80語（リストの全項目数は1865）で，全体に占める総出現頻度は5422回（リスト全体で10763回）であり，20回以上出現する「語」の全体リストに占める比率は約50％である。

3．試験テキスト，ウェッブ英文テキストとBNCとの比較

小さな英文テキストに現れる言語現象が，一般的な現象かデータに固有の現象であるか検証するためには，テキストデータが質量面でバランスの取れた大規模コーパスデータの広く知られている傾向と比較すべきである。ここでは，3つの試験問題とウェッブ上の英文テキストの分析結果を，The British National Corpus（以下，BNC）の研究成果と比較検討する。

なお，BNC の1億語に及ぶ英文テキストコーパスは，1985年から1994年にかけて当時の「書き言葉」と「話し言葉」データから収集された。現在，BNC はそのデータ量や分野間のバランスなどの面で，最も均衡のとれた大規模英文テキストコーパスとして知られている。

3．1　「語」の出現頻度リストの比較

3つの英語試験テキストの「リスニング問題」で出現頻度が高かい think と know について，BNC コーパスで頻度を調べてみると，BNC でも同一の傾向となった。（**データ3**を参照。ただし，表の頻度数値は100万語に対する相対頻度を示す。）

データ3　BNC データ1

語	品　詞	口語体	LL＊	文章体
know	動　詞	5550	104930.3	734
think	動　詞	3977	71946	562

＊LL Log Likelihood

一方，「リスニング以外の問題」で出現頻度が高い used について BNC コーパスで確認すると，動詞 used の用例は文章体が優勢であったが，準助動詞としての用法は口語体で顕著であった。（**データ4**参照。ただし，表の頻度数値は100万語に対する相対頻度を示す。）

データ4　BNC データ2

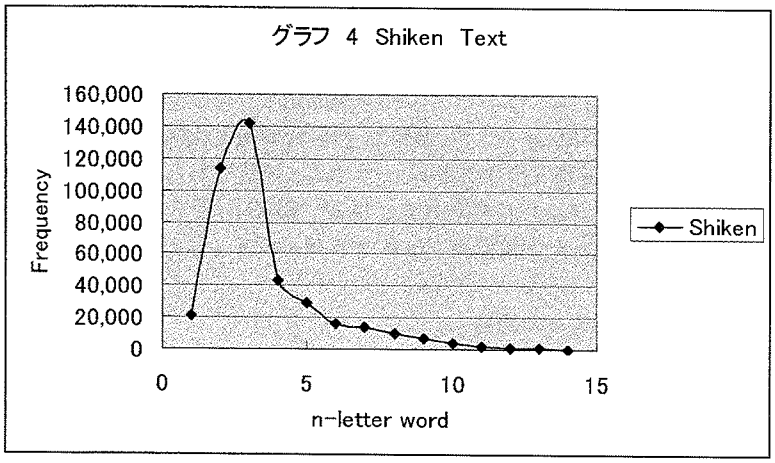
語	品　詞	口語体	LL	文章体
used	準助動詞	742	15000.9	88
used	動　　詞	225	-1783.4	497

参照のために，「リスニング問題」の used の用例を見ると，全18例の中で準助動詞の用法は1例で，それ以外は動詞としての用法であった。

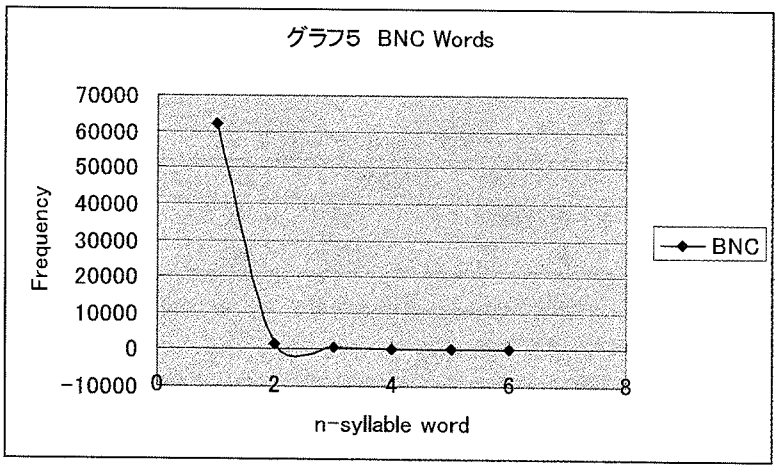
used については，3つの試験英語テキストと BNC では異なった傾向を示す結果となった。3つの試験問題テキストでは動詞としての used が優勢という傾向は，テキスト本来の特徴と考えられる。しかし，対象となったテキストのデータ量を考慮すると，偶然の結果という仮定も棄却できない。

３．２　N 文字語の分布状況

すでに見たように，「語」の文字数による頻度数を14段階（１文字語から14文字以上の「語」）に分けてグラフ化すると，３文字語あたりに出現頻度のピークが見られ，その後は右下がりとなって減少する。この頻度分布のパターンは，**Appendix 1** や **2** のテキストデータよりも大規模なデータでも認められる。（**グラフ 4** を参照。）ちなみに，**グラフ 4** のもととなった英文テキストは，TOEFL（５回分），TOEIC（５回分），英検１級・準１級・２級試験問題（それぞれ６回分），大学入試センター試験英語（1990年から2000年の本試験問題10回分）からなる英語試験問題で，そのデータ量は **Appendix 1** の全データの約２倍である。[3)]



テキストに現れる「語」の出現傾向について，Leech et al.(2001:121)は BNC データの「語」の長さを音節単位で比べている。音節を単位に「語」の分布を見ると，１音節語がピークとなり，２音節語(1,634)，３音節語(622)，４音節語(386)，５音節語(221)，６音節語(93)と右下がりに減少している。（括弧の中の数字は100万語あたりの語数。）



「語」の出現頻度を文字数で表示するか，音節数によるかの違いはあるが，**グラフ4**と**グラフ5**には数値のピークから急激な右下がり，その後は緩やかな減少という共通のパターンが認められる。

グラフ2～**グラフ5**の数値分布から，英文テキストに現れる「語」の最多は平均すると3文字から4文字[4]，あるいは1音節語であり，それ以外の文字数あるいは音節数からなる「語」の出現頻度は大変低い数値になることがわかる。なお，音節にもとづく**グラフ5**には，**グラフ4**などの1文字語から3文字語にわたる急激な上昇パターンが欠落している。これは，1音節語には，1文字語(例 I)，2文字語(例 so)，3文字語(例 sit)，4文字語(例 silk)などとともに，1文字1音とならない「語」(例 sketchや through)などの存在が含まれていることを考慮すると当然の現象と思われる。

3．3　JACET 基礎語彙8,000による比較

3つの試験テキストとウェッブ英語の語彙比較として，使用頻度の高い「語」を除いた後の両グループの語数を比べる方法がある。今回は，英語学習者にとって使用頻度の高い語彙リストとして「JACET 基礎語彙表」[5]の約8000語を使った。WordSmith にこの約8000語を Stop List として登録し，両グループの8000語を超える「語」の頻度順リストとその統計データを出力した。**データ5**は，Stop List の設定有無による「語」(Type) の数値を比較したもの。各テキストから8000語を除くと，語数は該当語を除く前の約1/2となっている。

データ5　Types の数値

	Stop List	
英文テキスト	指定なし	JACET8000
リスニング問題	4192	1872
リスニング問題以外	6343	3212
N. Korean Abductors	1865	843

ただし，8000語はレマリストであり，例えば，be 動詞 (am, is, are, was 等) はすべて集約されて be になっているのに対して，今回出力された語リストはレマ化されていないため，上記のそれぞれの語形は別の「語」としてカウントされている。また，今回の頻度順リストには，国名・地名・人名などの多くの固有名詞が含まれている。よって，**データ5**の JACET 8000 (8000語超の欄) の数値は，実際はもっと小さいものと推測される。例えば，ウェッブ英語(N. Korean Abductors)で見ると，be 動詞 (was(115/9236)；is(56/9982)；were(53/3227))，北朝鮮拉致事件関連の固有名詞 (Korea(147/19)，Koizumi(73/0)，Pyongyang(41/0)，abducted(26/0)，kidnapped(23/0)) などが高頻度語である。(括弧内の数字は，それぞれウェッブ英語/ BNC の頻度。)

このテキストでは頻度5以上の「語」は128語で，その多くは機能語または上記のようなテキスト関連の固有名詞であり，頻度1の「語」が380語にのぼることを考慮すると，このテキストの中で英語学習者が注目すべき8000語レベルを超える「語」はほとんど見あたらない。同様に，「リスニング問題」「リスニング以外の問題」についても，8000語レベルを超えかつ出現頻度の高い重要語はなかった。ちなみに，前者で頻度1の「語」は1075語，後者では958語であった。

これまでの観察を総合すると，固有名詞を除いた3つの試験テキストとウェッブ英語の高頻度「語」の多くは，JACET 8000語でカバーされていると考えられる。

4．おわりにかえて

英語の試験問題，あるいはウェッブでよく見る英文テキストに現れる「語」を比較検討するために，WordSmithによる解析を試みた。解析データをBNCコーパスの研究成果に照らし合わせることで，観察できた言語現象が通常の英文テキストに共通の傾向を示しているのか，個々のテキストに固有のものか検討するとともにその手順を明らかにした。英文テキストの中の「語」の出現傾向とその解析手順は，今後，さまざまなテキストを比較する場合に応用可能である。

コーパス言語学の視点から，英文テキストの難易度を「JACET 基礎語彙表」で検証する試みは，この語彙表の改訂4版リストの確定作業と共にさらに深めていきたい。

注

1）1999年～2001年の3カ年に渡るシーボルト大学学生の外国語学習実態調査の調査結果は学内LAN上にあり，学内者のみ閲覧可能。ただし1999年度と2000年度実態調査については，『国際情報学部紀要』(2000，2001)と『県立長崎シーボルト大学「共同教育研究費」に関わる研究報告書』(2001)に，中間報告として掲載されている。

2）本学国際交流学科ホームページの日本語ページ(http://www.mce.sun.ac.jp./index__j.html)と英語ページ(http://www.mce.sun.ac.jp./eng/index__e.html)を参照。

3）「JACET 基礎語彙表」改訂作業のために，本稿執筆者が責任者となって集めた英文コーパスデータの中の「試験英語テキストデータ」を使用。

4）「語」の平均文字数については，テキスト間でばらつきがある。**Appendix 1** と **2** のテキストでは，最小値はSTEP1（3.53文字），最大値はNYT（5.05文字）であつた。

5）「JACET 基礎語彙表」（第3版）は，英和辞典の見出し語の重要度ランクづけや，英語学習のための語彙表の作成基準として，10年近く活用されてきた。今回，WordSmithの Stop List として使った約8000語は，現在，改訂進行中の改訂4版の最終候補語。

参照文献

上村俊彦・他編(2001)「シーボルト大学学生の外国語学習実態調査報告(1)－1999年度実施調査の中間報告－」県立長崎シーボルト大学　「共同教育研究費」に関わる研究報告書 pp.1-14

Hunston, S. (2002) *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Leech, G,; Rayson, P. & Wilson, A (2001) *Word Frequencies in Written and Spoken English.* London: Pearson Education Limited.

Meyer, C. (2002) *English Corpus Linguistics: An Introduction.* Cambridge: Cambridge University Press.

Oxford University Computing Services(1999)*British National Corpus Sampler.* Oxford: Oxford University Computing Services.

テキスト

1．英語試験問題テキストファイル

1．1　**STEP1**

英語検定協会編　英検1級試験問題集（1999年前期・後期，1998年前期の3回分）

1．2　**TOEIC**

全国大学生協協同組合連合編(1980) TOEIC 第2回公開テスト問題

Lougheed, L. (1999) "TOEIC Model Test 4" in *Barron's How to Prepare for the TOEIC Test 2nd ed.*

The Chauncey Group International (2000) TOEIC公式ガイド& 問題集 「TOEIC 練習テスト」東京：(財)国際ビジネスコミュニケーション協会　TOEIC運営委員会

1．3　**TOEFL**

ETS (1998a) "Practice Test A" in *TOEFL Practice Tests Workbook Volume 1.*
ETS (1998b) "Practice Test B" in *TOEFL Practice Tests Workbook Volume 1.*
ETS (1998c) "Practice Test A" in *TOEFL Test Preparation Kit Workbook.*

2．ウェッブテキストファイル（下記データの URL は，すべて９月22日現在のもの。）

2．1　日本のウェッブサイト

Asahi (Asahi.com 朝日新聞社公式サイト)
Koizumi: I am deeply shocked by what I've heard and I strongly protest.
(http://www.asahi.com/english/politics/K2002091800528.html)
Koizumi off to N. Korea for summit with Kim.
(http://www.asahi.com/english/politics/K2002091700282.html)
JT (The Japan Times 公式サイト)
Kim admits abductions. Four Japanese alive, six dead; normalization talks to resume.
By JUNKO TAKAHASHI
(http://www.japantimes.co.jp/cgi-bin/getarticle.pl5?nn20020918a1.htm)
Abductees' families express grief, rage over death reports.
(http://www.japantimes.co.jp/cgi-bin/getarticle.pl5?nn20020918a2.htm)

2．2　米国ウェッブサイト

CNN (CNN 放送局公式サイト)
The missing Japanese. September 17, 2002 Posted: 8:36 AM EDT (1236 GMT)
(http://www.cnn.com/2002/WORLD/asiapcf/east/09/17/japan.kidnap.reut/index.html)
N. Korea admits Japanese kidnappings. September 17, 2002 Posted: 8:44 AM EDT (1244 GMT)
(*http://www.cnn.com/2002/WORLD/asiapcf/east/09/17/nkorea.japan/index.html*)
NYT (The New York Times 紙公式サイト)
North Koreans sign agreement with Japanese. By Howard W. French. *September* 18, 2002
(http://www.nytimes.com/2002/09/18/international/asia/18KORE.html?pagewanted=print&position=top)
WPost (The Washington Post 紙公式サイト)
North Korea admits abducting Japanese. Kim Jong Il pledges to freeze missile tests; Japan to restore diplomatic ties. Washington Post Foreign Service. Tuesday, September 17, 2002; 3:10 PM. (*http://www.washingtonpost.com/wp-dyn/articles/A28233-2002Sep17.html*)
For many families, hope dies. N. Korea confirms worst for Japanese whose loved ones disappeared. By Sachiko Sakamaki. Wednesday, September 18, 2002; Page A18
(http://www.washingtonpost.com/wp-dyn/articles/A31249-2002Sep17.html)
N. Korea admits it abducted Japanese: Disclosure clears way for historic accord. By Doug Struck. Washington Post Foreign Service. Wednesday, September 18, 2002; Page A01
(http://www.washingtonpost.com/wp-dyn/articles/A31387-2002Sep17.html)

２．３　英国ウェッブサイト

BBC（BBC 放送局公式サイト）

N Korea confesses to kidnappings. Tuesday, 17 September, 2002, 10:55 GMT 11:55 UK (http://news.bbc.co.uk/2/hi/world/asia-pacific/2262074.stm)

Analysis: Pyongyang's U-turn on abductions. By Charles Scanlon. Tuesday, 17 September, 2002, 18:00 GMT 19:00 UK. (http://news.bbc.co.uk/2/hi/world/asia-pacific/2264461.stm)

Times

(The Times 紙公式サイト http://www.timesonline.co.uk/ から Japanese abduction でサーチ)

North Korea admits kidnapping Japanese. By Caroline Gluck in Seoul. September 17, 2002

North Korea admits kidnap plot. From Caroline Gluck in Seoul. September 18, 2002

Appendices

Appendix 1: STEP, TOEIC & TOEFL Compared

	Listening Part of the Texts				Non-listening Part of the Texts			
Text File	OVERALL	SSTEP1	STOEIC	STOEFL	OVERALL	WSTEP1	WTOEIC	WTOEFL
Bytes	272,906	103,089	76,889	92,928	324,643	134,096	75,799	114,748
Tokens	48,308	20,242	13,888	14,178	54,412	25,976	12,145	16,291
Types	4,192	2,123	2,039	2,032	6,343	3,315	2,546	3,018
Type/Token Ratio	8.68	10.49	14.68	14.33	11.66	12.76	20.96	18.53
Sd. T/T R	41.42	45.29	40.07	39.98	46.22	52.3	45.69	42.11
Ave. Word	3.82	3.53	3.9	4.16	4.14	3.7	4.5	4.59
Sentences	4,644	882	2,368	1,394	2,587	667	1,000	920
Sent. Length	10.2	22.86	5.85	9.58	19.18	38.78	8.87	16.17
Sd. Sent. Length	10.32	14.63	4.03	7.63	17.6	19.77	7.94	10.78
1-letter words	3,819	919	1,690	1,210	4,253	1,031	1,523	1,699
2-letter words	10,427	5,576	2,502	2,349	12,044	7,544	1,786	2,714
3-letter words	12,928	7,188	2,798	2,942	13,664	8,745	2,271	2,648
4-letter words	7,584	2,272	2,595	2,717	6,502	2,534	1,713	2,255
5-letter words	4,684	1,643	1,438	1,603	4,142	1,681	981	1,480
6-letter words	2,702	701	942	1,059	3,391	1,026	917	1,448
7-letter words	2,399	680	792	927	3,418	1,150	905	1,363
8-letter words	1,607	516	476	615	2,554	881	716	957
9-letter words	1,018	329	336	353	1,975	583	650	742
10-letter words	560	220	145	195	1,227	386	335	506
11-letter words	284	91	75	118	652	231	163	258
12-letter words	163	59	51	53	309	109	89	111
13-letter words	102	35	38	29	164	47	54	63
14(+)-letter words	26	11	8	7	85	16	31	38

Appendix 2: Web Articles: North Korean Abductors

		日本サイト		米国サイト			英国サイト	
Text File	OVERALL	Asahi	JT	CNN	NYT	WPost	BBC	Times
Bytes	70,724	4,730	9,984	8,488	8,409	24,625	6,746	7,742
Tokens	11,451	762	1,600	1,386	1,330	4,018	1,095	1,260
Types	1,865	318	584	527	557	900	430	390
Type/Token Ratio	16.29	41.73	36.5	38.02	41.88	22.4	39.27	30.95
Sd. T/T R	44	NA	44.2	42.1	46.6	48.3	42.3	31.9
Ave. Word Length	4.88	4.89	4.94	4.8	5.05	4.9	4.77	4.83
Sentences	515	38	60	65	73	181	53	45
Sent. Length	21.67	18.74	26.18	19.97	18.22	21.78	20.58	27

Sd. Sent. Length	12.41	10.83	13.38	11.94	12.28	11.37	11.45	15.22
Paragraphs	312	24	49	43	30	86	50	30
Para. Length	36.49	31.75	32.65	32.23	44.33	46.3	21.9	41
Sd. Para. Length	21.06	16.86	16.21	15.98	14.48	25.45	14.11	21.77
1-letter words	377	37	55	55	34	129	27	40
2-letter words	1,735	122	238	208	215	540	199	213
3-letter words	2,075	144	305	236	223	774	195	198
4-letter words	1,807	96	216	247	201	664	184	199
5-letter words	1,585	105	215	199	179	546	147	194
6-letter words	914	44	130	107	113	339	70	111
7-letter words	940	75	136	100	120	327	95	87
8-letter words	776	44	117	93	76	297	63	86
9-letter words	618	45	102	69	71	195	60	76
10-letter words	265	19	43	33	35	83	26	26
11-letter words	171	16	21	25	25	58	10	16
12-letter words	85	4	12	6	12	34	9	8
13-letter words	72	8	9	7	15	23	5	5
14(+)-letter words	28	2	1	1	10	9	4	1

NA. WordSmith による該当出力データなし。

Appendix 3: STEP, TOEIC & TOEFL Compared (over JACET 8000)

	Listening Part of the Texts				Non-listening Part of the Texts			
Text File	OVERALL	SSTEP1	STOEIC	STOEFL	OVERALL	WSTEP1	WTOEIC	WTOEFL
Bytes	272,906	103,089	76,889	92,928	324,643	134,096	75,799	114,748
Tokens	48,308	20,242	13,888	14,178	54,412	25,976	12,145	16,291
Types	1,872	807	777	794	3,212	1,440	1,106	1,386

Appendix 4: Web Article: North Korean Abductors (over JACET 8000)

		日本サイト		米国サイト			英国サイト	
Text File	OVERALL	Asahi	JT	CNN	NYT	WPost	BBC	Times
Bytes	70,724	4,730	9,984	8,488	8,409	24,625	6,746	7,742
Tokens	11,451	762	1,600	1,386	1,330	4,018	1,095	1,260
Types	843	116	245	234	196	368	160	158