論　文

日本大学生産工学部研究報告B
2004 年 6 月 第 37 巻

日英パラレルコーパスを構成する テキストの難易度分類に関する研究

中條清美*・白井篤義**・内山将夫***・
西垣知佳子****・長谷川修治*****

A Study on Classifying Texts in English-Japanese Parallel Corpora According to Linguistic Difficulty

Kiyomi CHUJO, Atsuyoshi SHIRAI**, Masao UTIYAMA***, Chikako NISHIGAKI**** and Shuji HASEGAWA******

One of the most widely acknowledged barriers in utilizing corpora in the classroom environment is the lack of control over the concordance examples retrieved. This study focuses primarily on methods for classifying English or Japanese text sources in two English-Japanese parallel corpora based on their linguistic difficulty level. The study's purpose is to produce adequate learning materials for learners of English or Japanese as a foreign language.

In this study, four indices were applied to measure the linguistic difficulty of both English and Japanese text sample : 1) the readability of English texts ; 2) the text coverage not covered by JSH textbook vocabulary ; 3) the ratio of Levels 1 & 2 vocabulary in the Japanese language "Test Content Specifications" ; and 4) the ratio of kanji. It was found that each index provides sufficient criteria for classifying texts, that narrative texts are relatively easier than expository ones, and that the readability scores of English texts correlate highly with the other three indices. The authors believe the result of this research provides valuable information for promoting an efficient teaching method that makes the most of parallel corpora in English and Japanese language teaching.

キーワード：パラレルコーパス，テキスト分類，英語教育，日本語教育，難易度

1．はじめに

近年，英語教育分野でのコーパスの利用が急速に進みつつある。コーパスを使った指導法では，学習者自らが目的とする語句（キーワード）を検索してコンコーダンスと呼ばれる，キーワードの前後の文脈を表示するKWIC（Key Word In Context）検索の出力結果を使い，言語の実際使用を多数観察することによって，学習者自らが文法規則や語彙の意味・用法等を発見し学習するという帰納的な学習方法（DDL : Data-driven Learning）[1)]が主として使われる。

しかしながら，コーパスは本来「言語研究に使用されることを想定して，実際に書かれたり話されたりした言

*日本大学生産工学部教養・基礎科学系
**日本大学生産工学部応用分子化学科学部生
***情報通信研究機構
****千葉大学教育学部
*****千葉県立長狭高等学校

語をコンピュータ上で利用可能にしたテキストの集合体」[2]である。そのため，研究用に収集された資料をそのまま教育現場で利用しようとすると，「コンコーダンスの英語が難しすぎる」ために教材として機能しないという問題が生ずる[3]。学習者の英語力とコーパスの英語レベルとの間に大きなギャップが存在することから，コーパスの教育への応用利用を妨げている要因のひとつはコーパスを構成するテキストの難易度にあると考えられる[4]。

一般的に言語研究用にコーパスが作成される時には，収集されたテキストが研究対象となる言語のさまざまな変種（variety）を代表しているという代表性（representativeness）が重視される[5]。そのため，収集されるテキストの分野や量には関心が払われるが，テキストの英語の難易度に言及されることは多くないようである。このようにコーパスは元来教育への応用利用を考えて作られていない[注1]。一方，教材としてのコーパスの魅力は，言語の実際使用にもとづいた用例の「真正性（authenticity）」[6]にあり，それが学習者に対して説得力をもち，英語学習のやる気を喚起することに役立つ[7]。しかし同時に「真正性」は，本物であるがゆえにその英語レベルは高くなるという問題を抱えることになる。

このような状況を背景に，我々は，Utiyama（2003）[8]，内山他（2003）[9]により公開された日英パラレルコーパスに注目した。パラレルコーパスは同じ内容を2つの言語で記述したもので2言語コーパスとも呼ばれる[10]。**図1**は，“downtown”というキーワードを内山他（2003）の「読売新聞」と *The Daily Yomiuri* のパラレルコーパスより検索したコンコーダンス出力をCSVファイルに保存し加工したものである[注2]。**図1**の例のように，日英パラレルコーパスの場合は英語と日本語のデータを併せ持つため，たとえコンコーダンスの出力結果の英語が難しくても，並行して表示される日本語出力が補助情報となって英文を理解する負荷が大きく軽減されることになる[11]。

一般に辞書ではdowntownの訳語として「中心部」を第1義に載せており[12]，授業でもそのように指導している。しかし，downtownという語を学生はdownとtownの結合から類推して「下町」と訳す傾向があり，いったん日本語に置き換えると日本語の「下町」を連想してしまい，「上野，浅草」に代表されるような地域を意味する概念にずれてしまう。教室での指導においては当然，英語母語話者が使うdowntownの意味・概念を導くように指導しなければならない。その際，**図1**に示した具体例を多数示すことによって，学習者の「気づき（awareness）」[13]を促し，「繁華街，都心，市内，中心部」という実際の英語コミュニケーションにおける“downtown”の中心的な概念が“downtown”という語とネットワークで結ばれ易くなる。

図1の例のようにパラレルコーパスでは真正性を保った実際的言語データを英語と日本語の両方で提供できるので，英語だけのコンコーダンスより格段に言語素材の難易度を下げることが期待でき，教育上の利用価値は非常に高い。しかしながら，たとえ日本語を並行出力させ，内容理解の難易度をいくぶん下げることができたとしても，依然として，コーパスの英語そのものは高いレベルにあるため，学習者の習熟度レベルに見合った言語素材を提供するという効率的な学習に必須の教授要件を満たしているとは言いがたい。この問題を解決するには，あらかじめ何らかの基準でコーパスを構成するテキストの難易度を調査し，テキスト単位で難易度レベルに関する情報を蓄積し，それぞれの学習者が各自に適した言語レ

g on trees along streets and rooftops in	**downtown**	Tokyo. Conventionally orbiting satellites
繁華街の街路樹や建物の屋上にも営巣		するようになった。
big cities like Tokyo, some families living	**downtown**	kept roosters as pets. Another salient
ゴチャゴチャした**下町**のどこかの家が		ペットに鶏を飼っていて研究開発や情報
rconcentration of population. Flooding at	**downtown**	subway stations also halted traffic along
都心の地下鉄も浸水が相次ぎ，一部区		間で不通となった。その後，注文から決済
akes about 40 minutes by bus to get to	**downtown**	Abashiri from Memanbetsu Airport. The
◆あし　女満別空港から網走**市内**まで		バスで約40分。来年の通常国会に健康
ed two computers from gang hideouts in	**downtown**	Osaka and Ichikawa, Chiba Prefecture.
大阪・**ミナミ**と千葉県市川市のアジトから		パソコン一台ずつを押収。
ion plan to be finalized next spring. The	**downtown**	area became an enormous outdoor stage
大きな“劇場”と化した**市街地**は県内外		から集まった寺山ファンなどでにぎわった。
industry include the “hallowing out” of	**downtown**	areas in provincial cities and an increase I
急速に進む中で，地方都市**中心部**の空		洞化や中小小売店の廃業増加などの問題
uld be made to bring residents back to	**downtown**	areas. He was a football player and weight
都心居住を推進したい。		海軍兵学校時代はフットボールと重量挙げ

図1　パラレルコーパスの検索結果の例

ベルのテキストを選択できるような教授用のシステムが必要である。

我々は上述の利点を備えた日英パラレルコーパスを用いた上で，さらに学習者のレベルに合わせてテキストの難易度をコントロールしたいと考える。本研究で対象とする日英パラレルコーパスは多様なテキストから構成され，それらのテキストの英語の難易度，日本語の難易度，そして英語と日本語の対訳の難易度はどのような関係になっているかはまだ明らかにされていない。本研究では日英パラレルコーパスを構成するテキストを，学習者の英語習熟度レベルの観点から分類し，今後の日英パラレルコーパスを使った外国語教授法確立のための基礎的な資料としたいと考えている。

2．研究の目的

本研究の目的は，教育的視点に立った客観的な指標を用いて日英パラレルコーパスを構成している英語テキストおよび日本語テキストの難易度を計測することによって，難易度によるテキスト分類の可能性を確認することである。本研究の結果は，日英パラレルコーパスを活用した効果的な外国語教授法を構築する際の基礎的研究資料となるものである。

具体的には，現在公開されている利用可能な2種類の日英パラレルコーパスのテキスト33編について，英語テキスト部分と日本語テキスト部分を別個に分析した。英語テキスト部分は，1）英語リーダビリティ公式によるリーダビリティ・スコア，2）日本の中学・高等学校英語教科書に出現しない語の割合，また，日本語テキスト部分については，3）日本語能力試験の語彙1，2級の割合，4）漢字含有率，というテキストの内容理解に直接影響する要因を測る4つの指標を用いてテキストの難易度を調査した。

以下では，まず3節で本研究に使用した日英パラレルコーパスの特徴と，その中から調査に用いたサンプルテキストの作成方法について述べる。次に4節で，日本人英語学習者の英語習熟度と比較するため，使用した学校英語教科書の調査に用いたサンプルテキストおよび教科書語彙リストの作成方法について説明する。そして，5節で4種類の指標による難易度の測定方法を述べた後，6節で調査結果の考察を行ない，7節で計測のまとめを行なう。

3．日英パラレルコーパス

調査には，現在利用可能な日英パラレルコーパスとして，Utiyama(2003)[14]による①「日英対訳文対応付けデータ」（以下，散文データ）と内山他(2003)[15]による②「日英新聞記事対応付けデータ」（以下，新聞データ）を使用した。②の新聞データは約19万文よりなる世界最大規模の日英パラレルコーパスである。**表1**に，本研究で使用した2種類の日英パラレルコーパスを構成するテキストのうちサンプルとして選定した33編のタイトル，著者，ジャンル，延べ語数，日英対訳対の数を示した[注3)]。No. 1～31が①，No. 32～33が②を典拠とする。以下に，これら2種のパラレルコーパスの概要と調査のためのサンプルテキストの作成方法について述べる。

3.1　日英対訳文対応付けデータ（散文データ：表1のNo.1~31）

日英対訳文対応が付けられた古典的な名作や最近のコンピュータ関連のエッセイなど合計66編の作品（2003年12月現在）からなる。本研究ではそのうち31編を調査した。英語テキストは，「Project Gutenberg」[16]などで配布されている著作権が消滅した作品や，GNUプロジェクト[17]関連のドキュメントなど，再配布等が許可されている文書で構成されている。日本語テキストは，それら英語テキストの翻訳を再配布等が可能という条件下で公開している「プロジェクト杉田玄白」[18]正式参加作品を中心として収集されたものである。これらの英語テキストと日本語テキストの個々の文は人手により対応付けられ，

http://www2.nict.go.jp/jt/a132/members/mutiyama/align/index.html)

から対訳データ全体がダウンロードできる。

3.2　日英新聞記事対応付けデータ(新聞データ：表1のNo.32~33)

日英新聞記事対応付けデータは，「読売新聞記事データ」の1989年9月から2001年12月までの「読売新聞」と *The Daily Yomiuri* 記事を対応付けしたものである。新聞記事データの特徴は，ある程度翻訳の品質が保証され，かつ，分野やスタイルが均質なデータが大規模に得られる点にある。これら新聞記事の対応付けは自動的に行なわれ，対応付けデータは一般に公開されている（http://www2.nict.go.jp/jt/a132/members/mutiyama/jea/index.html）。

3.3　調査に用いたサンプルテキストの作成

上記の散文データ，新聞データの2種類のパラレルコーパスから次の方法を用いてサンプルテキストを作成した。散文データについては，Webページの日英対訳文対応付けデータリストから31編を選択し，それぞれ英語と日本語のテキストファイルを作成し保存した[注4)]。新聞データについては，Web上で公開されている日英対になった2つのサンプルテキストを取り出し，日本語部分と英語部分を分離して英語と日本語のテキストファイルを各1種類ずつ対になるよう作成した。

これらのテキストファイルは**表1**の延べ語数や対訳対の数からわかるように，サイズが多様であるため基準を設けてサンプルサイズを決定する必要がある。そこで，

表1　本研究で使用した日英パラレルコーパスの構成テキスト

No.	タイトル	著　者	ジャンル	延べ語数 英語	延べ語数 日本語	日英対訳対の数
1	The Adventure of the Norwood Builder	Conan Doyle	小説	9246	9779	810
2	The Declaration of Independence	United States	政治文	1335	1820	64
3	A Biographical Sketch of an Infant	Charles Darwin	記録文	4418	5880	246
4	A Dog of Flanders	Ouida	物語	14063	22482	1064
5	A Harlem Tragedy	O Henry	小説	2143	3094	197
6	A Midsummer Night's Dream（retold）	Charles and Mary Lamb	物語	4177	5494	331
7	A Scandal in Bohemia	Conan Doyle	小説	8523	9147	793
8	A Sense of History : Some Components	G.W. Schlabach	エッセイ	2200	3537	189
9	After Twenty Years	O Henry	小説	1271	1674	127
10	As You Like It（retold）	Mary Lamb	物語	5983	7454	398
11	Boycott Amazon !	R.M. Stallman	投稿文	1023	1359	70
12	Clinton's Inaugural Address	Bill Clinton	政治演説	1606	2465	122
13	Dubliners	James Joyce	小説	1829	2325	166
14	Free Software is More Reliable !	Free Software Foundation	投稿文	560	834	47
15	Freedom-Or Copyright ?	Richard Stallman	投稿文	765	1312	60
16	Hearts and Hands	O Henry	小説	871	1382	104
17	JFK's Inaugural Address, 1/20/1961	J.F. Kennedy	政治演説	1428	1990	81
18	Linux and GNU-GNU Project	Richard Stallman	解説文	1103	1563	74
19	Peter Pan in Kensington Gardens	James M. Barrie	童話	15815	24045	1254
20	Romeo and Juliet（retold）	Charles and Mary Lamb	物語	6657	9512	403
21	Snow White and the Seven Dwarfs	Grimm brothers	童話	3110	4148	264
22	Some Confusing or Loaded Words and Phrases that are Worth Avoiding	Free Software Foundation	投稿文	1095	1675	93
23	Someone to Watch over Me	Richard Stallman	エッセイ	205	325	13
24	The Abolition of Work	Bob Black	学術論文	6661	10510	463
25	The Adventure of the Blue Carbuncle	Conan Doyle	小説	7832	9883	715
26	The Adventure of the Devil's Foot	Conan Doyle	小説	10016	12434	799
27	The Arrest of Arsene Lupin	Maurice Leblanc	小説	4875	6895	460
28	The Assignation	Edgar Allan Poe	小説	4605	6715	334
29	The Beginning of Ownership	Thorstein Veblen	学術論文	5189	5328	207
30	The Black Cat	Edgar Allan Poe	小説	3987	5898	224
31	The Darwinian Hypothesis	T.H. Huxley	学術論文	5243	5395	211
32	読売新聞/The Daily Yomiuri p11-sample		新聞記事	1626	2090	100
33	読売新聞/The Daily Yomiuri pnm-sample		新聞記事	2638	3349	100

1）散文データはストーリー性のあるものなので，可能な限り各テキスト全体を調査する，2）後述する日本語解析プログラムの処理能力に合わせる，という2つの要件を考慮して，日本語のテキスト部分が19 KB以下のものはテキスト全体を1サンプルとし，19 KB以上のものは2サンプル（1サンプルは日英200文対とする）をランダムに抽出することにした[注5)]。結果，散文データ31編，新聞データ2編，計33種類のテキストから日英のそれぞれについて各46サンプルテキストが作成された。

4．学校英語教科書

日本人英語学習者の読書能力と語彙力の目安を推定できる言語資料として，我が国で使用されている中学・高等学校英語教科書を用いた。

4.1　学校英語教科書

調査に使用した教科書は，中学校教科書は *New Horizon 1*，*2*，*3*（東京書籍，2000）[19]，高等学校教科書は *Unicorn I*，*II*，*Reading*（文英堂，1997，1998，1999）[20]である。一般的な傾向が得られるように，中学校，高等学校，ともに採択数の多い教科書シリーズとした[21]。高等

学校用教科書は，高校修了時レベルの上限の一例として，難易度の高い上級の教科書シリーズを使用した。

4.2　リーダビリティ調査用サンプルテキスト

日本人英語学習者の読書力の目安として，英語リーダビリティ公式を利用してリーダビリティを求めるため，上記の学校英語教科書より以下の英文テキスト部分を抜粋して，合計 15 サンプルを作成した。

① 中学教科書 *Horizon 1, 2, 3* より各学年各 2 箇所の“Let’s Read”

② 高校教科書 *Unicorn I*，*II*，*Reading* の各 Lesson 1，5，10

高校教科書は *I*，*II*，*Reading* の順に高等学校の 1，2，3 学年で学習すると仮定した。また，リーダビリティを正確に算出するには，英文サンプルテキストから固有名詞，数字，略語等を除外する必要があるので[注6]，人手によりそれらの語を除いた。

4.3　英語教科書未知語率調査用語彙リストの作成

英文テキストの難易度を測定するための指標のひとつとして，英語教科書に出現しない語の割合（未知語率）を算出した。語彙リストの作成にあたっては，上記の教科書の本文と Supplementary Reading の部分を入力し，校正，編集を施した後，固有名詞，数字，略語等を除いた。語彙リストの屈折形は基本形に集約した。

本稿では，日本人英語学習者の英語レベルを大学入学者レベル（高等学校修了レベル）に設定した。大学入学者は中学・高等学校教科書を使用して英語の主要部分を学習してきているため，中学・高等学校教科書で習得する語彙を学習者の語彙力レベルと仮定した。一人の学習者が使用する平均的な教科書の例として，上記の中学校教科書と高等学校教科書を組み合わせて学校英語教科書語彙リスト（異語数：3098 語，延べ語数：43772 語）を作成した。このリストを後述の 5.2 で学校英語教科書でカバーされていない語の割合（英語教科書未知語率）の算定に用いた。

5．調査方法

テキストの難易度を客観的に測定，あるいは推定できる指標として，**表 2** に示した 4 項目を調査した。

5.1　英語テキスト：英語リーダビリティ

英語リーダビリティは，「文章を読みやすくする要因，すなわち単語の難易，単語の長さ，センテンスの長さなどの要因を組み合わせ，公式に代入して計算し，その数字を読書学年レベルとするものである」と定義されている[22]。算出結果は，通常，米国の読書学年に相当する「学年」で表される。たとえば，リーダビリティ・スコアの「8.0」は米国の 8 年生の生徒がその文書・読み物を理解できることを意味する[注7]。米国では，最初のリーダビリティ公式が 1928 年に発表されて以来[23]，リーダビリティの指標が教育界に広く普及しており，リーディングや ESL（English as a Second Language）の教師に日常的に利用されている[24]。現在は数種類の公式が組み込まれたソフトウェアが利用可能で[25]，その実用性は高い。

本研究では，そのようなソフトウェアの 1 つである Readability Calculations[26]を利用し，**表 3** に示した 9 種類の公式を用いて，英語サンプルテキストのリーダビリ

表 2　テキストの難易度を客観的に測定するための調査項目

英語テキスト	日本語テキスト
・リーダビリティを示す推定値 ・学校英語教科書に出現しない語の割合	・日本語能力試験 1，2 級の語彙の構成比 ・漢字含有率

表 3　9 種類のリーダビリティ公式

		最大値（学年）	最小値（学年）	平均値（学年）	標準偏差	要　因	対　象（参考）
①	Dale-Chall Formula	9.6	5.0	6.7	1.2	語彙リスト，語数，文長	小学校高学年～中学校
②	Flesch Reading Ease Formula*	95.0	31.8	72.8	15.3	語数，シラブル数，文数	成人向け読み物
③	Flesch-Kincaid Formula	14.3	2.7	6.2	2.7	語数，シラブル数，文数	小学校高学年～中学校
④	FOG Formula	29.8	6.6	14.6	5.8	語数，3 シラブル以上の語数，文数	中学校以上
⑤	Powers-Summer-Kearl Formula	7.7	4.1	5.3	0.9	語数，シラブル数，文数	小学校低学年
⑥	SMOG Formula	15.1	5.8	9.1	2.1	3 シラブル以上の語数	4 年生～18 年生
⑦	FORCAST Formula	11.8	7.4	9.3	1.1	1 シラブルの語数	成人向け
⑧	Spache Formula	6.2	2.5	3.6	0.7	語彙リスト	小学校低学年
⑨	Fry Graph	17.0	2.5	7.1	3.6	シラブル数平均	小学校～大学
	(③+⑥+⑨)/3	15.1	3.7	7.5	2.8		

＊ ②の単位は学年でなく，0 ～100 のスコアで表示され，低スコアは高学年を意味する。

ティ・スコアを算出した。9種類の公式はリーダビリティ関連の文献に頻出の公式であるため，すべての公式によるスコアを算出してみる必要があると考えた。なお，同一テキストからサンプルテキストを2個抽出した場合にはそれらのスコアの平均を求めた。**表3**には33種類の英語サンプルテキストについてリーダビリティ公式によって算出された読書学年の最大値，最小値，平均値，スコアの標準偏差，リーダビリティ公式の算出に用いた要因，各公式に推奨されている最適対象学年を示した。

表3から明らかなように，算出スコアは公式によって大幅に異なる。当該調査資料にもっとも妥当な公式を選定する必要があるので[注8]，**表3**の「要因」と「対象（参考）」にあげた項目，および，9種類の公式を適用した読書学年レベルの算出結果を照合して検討した。たとえば，新聞記事や童話の *Snow White* の算出学年が適切な学年に配当されているかなど，算出されたリーダビリティの読書学年の妥当性，さらに，算出された読書学年の範囲が十分な幅をもって算定されているか等を吟味した。その結果，**表3**で網掛けで示した③ Flesch-Kincaid Formula[28]，⑥ SMOG Formula[29]，⑦ Fry Graph[30]の3指標が本稿の調査サンプルのリーダビリティの予測に適当と判断された。本稿ではこれら3指標で算出された読書学年の平均を利用することにする。他の6指標を使用しない理由は以下のとおりである。

1）Dale-Chall Formula[31]は正確な指標と報告されているが[32]，本研究の資料については表示の幅が5年～9年と狭かったので除外した。
2）Flesch Reading Ease Formula[33]は学年表示でなく0－100で表示される。
3）FOG Formula[34]は他の指標の2倍程度高いスコアを表示した。
4）Powers-Summer-Kearl Formula[35]とSpache Formula[36]は7年生以上は表示されなかった。
5）FORCAST Formula[37]は，表示幅が7年～11年と狭かったので除外した。

なお，日本人英語学習者の英語読書力の目安としてのリーダビリティを求めるために，4節で言及した中学・高等学校英語教科書の15サンプルの抜粋テキストについてもリーダビリティ・スコアを算出した。算出には上述したFlesch-Kincaid Formula, SMOG Formula, Fry Graphの3公式のスコア平均を学年ごとに求め，その結果を**図2**に示した。

それぞれの棒グラフに記されている数値は，米国における生徒の読書能力を学年で表示したもので，読書学年レベル（reading grade level）を予測する数値である[注9]。**図2**からは，ほぼ学年とともにレベルが上昇し，高校3年生の教科書は米国の中学3年～高校1年程度に相当する8.9年と推定された。本稿では，この8.9年生を日本人英語学習者の読書学年レベルと仮定して以下の考察をすすめる。

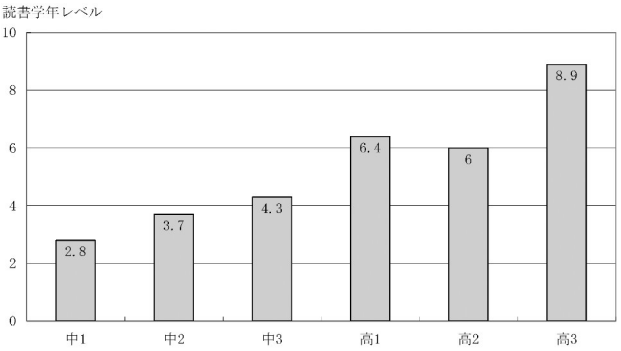

図2　中学・高等学校英語教科書のリーダビリティの推移

5.2　英語テキスト：学校英語教科書の未知語率

コーパスを構成する英文テキストと日本人英語学習者の語彙レベルとの関連を密接に知るために，各サンプルテキストの延べ語数のうち，中学・高校教科書に出現しない語数の割合を調査した。調査対象の英文テキストの語彙がどれくらい理解できるかを示す目安であるカバー率については研究が進んでおり，英文の内容理解に支障をきたさないためには最低限95％カバー率を満たす語彙数が必要といわれている[40]。本稿でも95％カバー率を内容理解に必要なカバー率の基準とする。その場合，英文テキストの出現語彙のうち，学校英語教科書使用語彙でカバーできない語の割合（未知語率）は5％以下でなければ，内容理解は難しいであろうということになる。

5.3　日本語テキスト：日本語能力試験の語彙1，2級の構成比

パラレルコーパスにおける日本語のテキストは，日本人英語学習者にとって，英語テキストの内容理解を支援する役割を果たす。また逆に，外国人英語母語話者の日本語学習者にとっては日本語の学習用教材となるため，各自の日本語レベルに合わせてテキストを選べるように難易度を調査する必要がある。日本語能力試験出題基準(2002)[41]には読解能力の測定対象について，「テキストに関して数量的に操作できるものは「語彙」「漢字」「文の長さ」といったものであり，内容的なものの難易を比較することはできない。」[42]と記されていることから，本研究では，「語彙」と「漢字」について調査することとし，Web上で公開されている日本語学習者のための読解学習支援システム“リーディング・チュウ太（Reading Tutor）”（http://language.tiu.ac.jp/tools.html）[43][44]の語彙チェッカーおよび漢字チェッカーを利用した。

語彙チェッカーは入力テキストをまず形態素解析システムの「茶筌2.02」を用いて解析し，その結果を日本語能力試験出題基準の1級から4級までの語彙リストと照合し，テキスト中の語彙の1級・2級・3級・4級・級外の級別分類表を作成する[45]。川村(2001)[46]では3級と4級の語彙含有率を合計して3級・4級の語彙含有率の

高い日本語テキストを「易しい」と判定している。本研究では，他の３つの指標との難易度の方向性を統一するため，この級別分類表を利用して，級外を除く１級から４級の語彙のうち，１級・２級の語彙の構成比を難易度のデータとした。従って，１級・２級の語彙の割合が高い方が「難しい」ことになる。テキストによってサンプルが２個の場合はレベル別語彙構成比の平均を求めた。

5.4　日本語テキスト：漢字含有率

日本語能力試験出題基準ではテキストに関する漢字の数量的基準[47]について，漢字含有率，すなわち「テキストの全字数（句読点を含まない）中の漢字の割合」[48]を設定している。本研究でも漢字含有率を調査することにした。まず，「漢字チェッカー」（川村 1999）[49]の出力結果からテキスト中の使用漢字数を求め，次に，Microsoft WORDの文字カウント（全角文字＋半角カタカナの数）を用いてテキストの全字数を求めた後，漢字含有率を計算した[注10]。

6．結果と考察

日英パラレルコーパスを構成する英語テキストと日本語テキストについて，４つの指標によって難易度を測定した結果を**表４**に示した。**表４**には，テキストの難易度の位置付けが把握しやすいように，学校英語教科書のリーダビリティ・スコアも含め，さらにテキスト全体をリーダビリティのスコアを基準にして昇順にソートした結果をそのテキストの属するジャンルとともに示した。

以下では，初めに，英語テキストのリーダビリティ，および学校英語教科書語彙の未知語率について観察し，次に，日本語テキストの日本語能力試験の語彙１，２級の構成比，および漢字含有率の比較，検討を行なう。最後に，４つの指標間の相関係数を求めた結果を考察する。

6.1　英語テキスト：リーダビリティ

英語テキストのリーダビリティ・スコアを**表４**の第５列に示した。また，日本人英語学習者のリーディング能力の目安と仮定した高校３年生用教科書の「8.9 年生」レベルを超えるスコアを網掛けで示した。33 編のテキストのうち９編（27%）は 8.9 年生を超え，残り 24 編（73%）は 8.9 年生以下である。リーダビリティの観点からは，今回調査した日英パラレルコーパスの構成テキストの７割以上の英文レベルは日本人英語学習者にほぼ適切なレベルと思われる[注11]。

宮浦（2002）[51]によると，英文読解の研究では，テキストは物語文（narrative）と説明文（expository）に二分される。**表４**で小説や童話など物語文に分類されるテキストの多くは，本稿で日本人英語学習者の読書学年レベルと仮定した 8.9 年生以下のものが多いことがわかる。宮浦（2002）[52]によると一般に物語文の読解は説明文に比べて易しい。その理由は，物語文で描かれる対象は日常生活での社会的関係であり，語彙もなじみのあるもので，物語の構造が比較的平易なので，読み手は物語スキーマの知識を活かすことができるためである[53]。

これに対し，**表４**において説明文に分類される新聞記事，学術論文，解説文，投稿文，演説原稿，エッセイ等は比較的リーダビリティが高く，多くは 10 年生レベル以上に算定されている。英文読解の研究においても，説明文の読解は比較的難しいとされている。その理由としては，題材が読み手の知識範囲を超えていることが多く，用語の意味のアクセスが困難を伴い，利用できる関連知識も少ない等さまざまな要因が考えられている[54]。リーダビリティの学年レベルが 10 年生以上と評価されているテキストは，本稿の共著者のうち英語教育に携わる３名の主観的な難易度評定でも難しいレベルと評価された。

6.2　英語テキスト：学校英語教科書の未知語率

英語テキストの使用語彙のうち中学・高等学校英語教科書語彙に出現しない語の割合を**表４**の第６列に示した。**表４**から明らかなように，内容理解の閾値といわれる未知語の占めるパーセンテージが５%以下をクリアしている英文テキストは調査対象の中になかった。この結果から，学校英語教科書の語彙だけを習得した日本人英語学習者がこれらの英語テキスト，あるいは，それらから抽出されるコンコーダンス出力に表示された文に対峙した折に経験するであろう困難さが容易に予測できる。学校英語教科書の語彙選定の妥当性については別稿で論じたい。

表４からは，リーダビリティの項の観察と同様，説明文は物語文より，一部の例外を除き，未知語の割合が多く難しいということも読み取れる。外国語としての英語教育における読解においてはとりわけ語彙の影響が大きいので，テキストを選択する際には注意が必要である[55]。コーパスを実際の指導に活用する前に学習者の語彙レベルを妥当な方法で測定し，その結果をもとに学習者の語彙レベルに近いテキストを自動的に検索して選定するプログラムをコーパスに組み込む必要があるかもしれない。

6.3　日本語テキスト：日本語能力試験の語彙１，２級の構成比

日本語テキストで使用されている日本語語彙のうち，日本語能力試験の出題基準に認定されている１，２級の語彙の割合を**表４**の第７列に示した。１，２級の語彙には，たとえば，「あし（足），あじ（味），アジア，あしあと（足跡），あしからず」などがある[56]。**表４**より１，２級の語彙の割合は物語文の場合，ほぼ10%台であり，説明文の場合は 20%を越え，新聞記事では 36%となっており，説明文では難しい日本語語彙が使われていることが判明した。

表4　英語テキストと日本語テキストの難易度測定の結果

No.	テキスト名	著者	ジャンル	英語リーダビリティ（読書学年）	英語教科書未知語率（%）	日本語の語彙 1＋2級（%）	漢字含有率（%）
	New Horizon 1 （中学1年）			2.8			
	New Horizon 2 （中学2年）			3.7			
21	Snow White and the Seven Dwarfs	Grimm brothers	童話	4.2	8.7	11.6	17.6
	New Horizon 3 （中学3年）			4.3			
6	A Midsummer Night's Dream（retold）	Charles and Mary Lamb	物語	4.8	10.6	12.8	17.2
16	Hearts and Hands	O Henry	小説	4.9	10.0	12.9	20.1
9	After Twenty Years	O Henry	小説	5.0	9.0	15.1	19.1
1	The Adventure of the Norwood Builder	Conan Doyle	小説	5.1	9.0	15.9	21.5
5	A Harlem Tragedy	O Henry	小説	5.1	12.3	17.6	18.9
19	Peter Pan in Kensington Gardens	James M. Barrie	童話	5.2	6.8	11.0	16.5
25	The Adventure of the Blue Carbuncle	Conan Doyle	小説	5.2	10.5	17.2	21.9
7	A Scandal in Bohemia	Conan Doyle	小説	5.3	9.6	17.8	26.6
4	A Dog of Flanders	Ouida	物語	5.4	12.6	14.9	20.3
13	Dubliners	James Joyce	小説	5.5	6.1	11.1	22.3
27	The Arrest of Arsene Lupin	Maurice Leblanc	小説	5.7	11.1	15.9	21.6
	Unicorn I （高校1年）			6.0			
10	As You Like It（retold）	Mary Lamb	物語	6.0	10.5	14.4	21.6
26	The Adventure of the Devil's Foot	Conan Doyle	小説	6.0	9.8	17.2	24.2
20	Romeo and Juliet（retold）	Charles and Mary Lamb	物語	6.3	12.7	15.4	21.1
	Unicorn II （高校2年）			6.4			
28	The Assignation	Edgar Allan Poe	小説	7.7	14.7	16.8	27.0
12	Clinton's Inaugural Address	Bill Clinton	政治演説	7.7	10.4	24.6	25.2
30	The Black Cat	Edgar Allan Poe	小説	7.9	18.1	15.9	23.2
23	Someone to watch over me	Richard Stallman	エッセイ	8.0	9.0	20.0	24.6
17	JFK's Inaugural Address, January 20, 1961	J.F. Kennedy	政治演説	8.0	11.1	24.8	27.9
3	A Biographical Sketch of an Infant	Charles Darwin	記録文	8.1	8.8	16.1	27.4
11	Boycott Amazon !	R.M. Stallman	投稿文	8.5	11.8	22.6	26.8
18	Linux and GNU - GNU Project	Richard Stallman	解説文	8.8	10.5	20.4	18.1
8	A Sense of History : Some Components	G.W. Schlabach	エッセイ	8.8	10.4	19.8	28.5
	Unicorn Reading （高校3年）			8.9			
22	Some Confusing or Loaded Words and Phrases that are Worth Avoiding	Free Software Foundation	投稿文	10.0	16.6	27.6	30.4
24	The Abolition of Work	Bob Black	学術論文	10.1	16.6	25.8	31.9
14	Free Software is More Reliable !	Free Software Foundation	投稿文	10.1	19.2	24.3	20.4
15	Freedom - Or Copyright ?	Richard Stallman	投稿文	10.2	16.0	24.4	28.7
31	The Darwinian Hypothesis	T.H. Huxley	学術論文	10.4	14.7	23.6	30.7
29	The Beginning of Ownership	Thorstein Veblen	学術論文	11.2	18.0	28.2	36.5
2	The Declaration of Independence	United States	政治文	11.6	17.6	27.5	34.2
32	読売新聞/The Daily Yomiuri p11-sample		新聞記事	14.3	16.1	36.1	44.3
33	読売新聞/The Daily Yomiuri pnm-sample		新聞記事	15.1	16.0	36.9	42.2
	平均			7.8*	12.3	19.9	25.4
				高3以降(9.0以降)			1，2級(25%以上)

＊ 学校英語教科書を含まない

日英パラレルコーパスは，日本人英語学習者が英語学習に利用するだけでなく，外国人英語母語話者の日本語教育にも応用できることが大きな利点である。実際に，日本語教育のさかんな豪州の西オーストラリア大学では本稿で調査対象とした「日英新聞記事対応付けデータ」を日本語上級クラスの翻訳コースの教材に使用するプロジェクトを開始した。この場合，英文のサポートによって日本語を理解する負荷が大いに軽減されることになる。従って，日英逆の立場から見た場合でも日本語テキストの難易度の把握は重要である。

表 5　テキストの難易度を測定する指標間の相関係数

	英語 リーダビリティ	英語教科書 未知語率	日本語の語彙 1＋2 級	漢字 含有率
英語リーダビリティ	—			
英語教科書未知語率	0.721	—		
日本語の語彙 1＋2 級	0.931	0.683	—	
漢字含有率	0.881	0.577	0.878	—

6.4　日本語テキスト：漢字含有率

非漢字圏の学習者の場合，あるいは日本人の学習者でも漢字習熟度の低い学習者の場合，漢字の負担を考慮する必要がある。日本語能力試験の出題基準には各級ごとに漢字含有率が，1 級 30－45%，2 級 25－35%，3 級 20－25%，4 級 15－20%と明示されている[57]。**表 4** には漢字含有率を第 8 列に示し，日本語能力試験の 1 級，2 級にあたるものを網掛けで示した。

他の 3 つの指標における観察と同様に，物語文の漢字含有率はほぼ 3 級か 4 級に相当する「易しい」ものであることがわかった。英語テキストのリーダビリティの 10 年生以上に算出された説明文の日本語テキストは漢字含有率が 1 級に相当し，新聞記事には非常に多くの漢字が使用されていることがわかる。

6.5　4 つの指標間の相関係数

以上の各指標の算出結果の考察を通じて，各指標は日英パラレルコーパスのテキストに対して類似した難易度評価を与えている傾向が観察された。そこで，4 指標間の相関係数を求め，その結果を**表 5** に示した。

表 5 より明らかなように，英語リーダビリティは，他の 3 つの指標，すなわち，英語教科書の未知語率，日本語の語彙 1＋2 級の割合，漢字含有率と高い相関があるという結果となった。なかでも，英語リーダビリティが英語教科書の未知語率よりも日本語の 2 指標との相関の方が高かったのは興味深い。

一方，日本語の語彙 1＋2 級の割合と漢字含有率は当然，高い相関があり，さらに英語教科書の未知語率も日本語の 2 つの指標とかなり相関があることが判明した。これらの結果は，今後，パラレルコーパスの英語テキスト，日本語テキスト選択の際の貴重な資料となると考える。

7．まとめ

コーパスの教育利用を困難にしている原因の 1 つにコーパスを構成するテキストの言語が難しすぎることがある。利用する学習者がコンコーダンスの出力結果を理解できるようになってはじめて，文法の規則や語彙の意味・用法の発見学習が可能となり，学習指導用の言語材料として機能する。従い，コーパスを教材として使用するには，学習者の習熟度レベルに合致したテキストを選択できるようにする必要がある。そのため，まず，コーパス構成テキストの難易度を明らかにする必要があった。

本研究では，日英パラレルコーパスを構成するテキスト 33 編について，英語テキスト部分については，1）英語リーダビリティ，2）学校英語教科書に出現しない語の割合，また，日本語テキスト部分については，3）日本語能力試験の語彙 1，2 級の構成比，4）漢字含有率，というテキストの理解に直接影響する要因を測定する 4 つの指標を用いてテキストの難易度を調査した。その結果，以下のことが導かれた。

- 4 つのいずれの指標を用いてもそれぞれの基準にもとづいてテキストの難易度別の分類が可能である。
- 物語文は説明文より一般的に易しい傾向があり，両者の難易度の差は大きい。
- 英語リーダビリティは他の 3 指標と相関が高く，日本語能力試験の語彙 1，2 級の構成比，漢字含有率の指標とも高い相関がある。

本研究の結果は，コーパスを外国語学習に利用する際，学習者のレベルや学習目的に適合した難易度のテキストをフィルタリングによって選択収集するための基礎資料のひとつになると考えられる。具体的には，コンコーダンスの出力結果の表示に関して，本稿で試みたような文脈を含むテキスト分類と，単語頻度によるセンテンスの難易度分類等[58][59]を組み合わせるとより広範でユーザーフレンドリーなコーパス利用につながると考える。

今回の調査に，文法項目は含まれていない。今後は文法項目によるテキストの難易度レベルの調査を行なうとともに，日本語のサポートによって英語の難易度がどの程度下がるのかという課題についても調査したいと考えている。

参考文献

1）Tim Johns, “Data-Driven Learning : the Perpetual Challenge”, *Proceedings of the Fourth Teaching and Learning Corpora (TALC) Conference*, Graz, 7/19-23/2000, http://www-gewi.kfunigraz.ac.at/talc2000/Htm/home.htm

2）齊藤俊雄，中村純作，赤野一郎，『英語コーパス言語学－基礎と実践』，研究社出版，東京，1998.
3）投野由紀夫，「コーパスを英語教育に生かす」，『英語コーパス研究』，第10号，2003，pp.249-264.
4）Wible, D., Feng-yi Chien, Chin-Hwa Kuo & Wang, T. C., "Adjusting Corpus Searches for Learners' Level － Filtering Results for Frequency", *Proceedings of the Fourth Teaching and Learning Corpora (TALC) Conference*, Graz, 7/19-23/2000, http://www-gewi.kfunigraz.ac.at/talc2000/Htm/home.htm
5）齊藤他，前掲書.
6）米山朝二，『英語教育指導法事典』，研究社，東京，2003.
7）Rixon, S., "Authenticity", In Johnson K. & Johnson H. (Eds.), *Encyclopedic Dictionary of Applied Linguistics : A Handbook for Language Teaching* Oxford : Blackwell Publishers Ltd., 1998, pp.68-69.
8）Utiyama, Masao., "Japanese-English Bilingual Corpora and their Applications", Asialex 2003. http://www2.crl.go.jp/jt/a132/members/mutiyama/publications.html
9）内山将夫，井佐原均，「日英新聞の記事および文を対応付けるための高信頼性尺度」，『自然言語処理』，第10巻，第4号，2003，pp.201-220.
10）金明哲，村上征勝，永田昌明，大津起夫，山西健司，『言語と心理の統計－ことばと行動の確率モデルによる分析』，岩波書店，東京，2003.
11）投野，前掲論文.
12）井上永幸，赤野一郎，『ウィズダム英和辞典』，三省堂，東京，2003.
13）石黒昭博，山内信幸，赤松信彦，北林利治，『現代の英語科教育法』，英宝社，東京，2003.
14）Utiyama（2003），前掲論文.
15）内山他（2003），前掲論文.
16）http://promo.net/pg/
17）http://www.gnu.org/
18）http://www.genpaku.org/
19）浅野博他，*New Horizon 1, 2, 3*，東京書籍，東京，2000.
20）末永國明他，*Unicorn I*，*II*，*Reading*，文英堂，東京，1997，1998，1999.
21）時事通信社，『内外教育』，1/16，1/30，2001.
22）高梨庸雄，卯城祐司，『英語リーディング事典』，研究社出版，東京，2000.
23）高梨・卯城，上掲書.
24）Fry, E. B., Kress J. E. & Fountoukidids, D. L., *The Reading Teacher's Book of Lists*, West Nyack, New York : The Center for Applied Research in Education, 1993.
25）Micro Power & Light Co., *Readability Calculations*, 2003.
26）Micro Power & Light Co., *Readability Calculations*, 2003.
27）中條清美，長谷川修治，「語彙のカバー率とリーダビリティから見た大学英語入試問題の難易度」2004予定.
28）Flesch R., *The Art of Readable Writing*. New York : Harper and Row, 1974. (as cited in Smith, C. R. & Smith, C. A. "Patient Education Information : Readability of Prosthetic Publications", *Journal of Prosthetics & Orthotics*, 6, 4, 1994, 113-118.)
29）McLaughlin, G., "SMOG Grading : A New Readability Formula", *Journal of Reading*, 12, 8, 1969, pp.639-646.
30）Fry, E. B., "A Readability Formula That Saves Time", *Journal of Reading*, 11, 7, 1968, pp.265-271.
31）Dale, E., & Chall, J. S. *A Formula for Predicting Readability*. Columbus, OH : Ohio State University Bureau of Educational Research, 1948.
32）高梨・卯城，上掲書.
33）Flesch, R., "A New Readability Yardstick", *Journal of Applied Psychology*, 32, 3, 1948, pp.221-233.
34）Gunning R., *The Technique of Clear Writing*. New York : McGraw-Hill, 1968. (as cited in Smith, C. R. & Smith, C. A. "Patient Education Information : Readability of Prosthetic Publications", *Journal of Prosthetics & Orthotics*, 6, 4, 1994, 113-118.)
35）Powers-Summer-Kearl Formula (as cited in Micro Power & Light Co., *Readability Calculations*, 2003.)
36）Spache, G., "A New Readability Formula for Primary Grade Reading Materials", *Elementary School Journal*, 55, 1953, pp.410-413.
37）Sticht, T. G., "Research toward the Design, Development, and Evaluation of a Job-functional Literacy Training Program for the United States Army", *Literacy Discussion*, 4, 1973, 339-369.
38）新井七菜子，「日本人英語学習者向けのリーダビリティ公式の提案と評価に関する研究」『第42回大学英語教育学会全国大会要綱』，2003，pp.95-96.
39）宮崎佳典，工藤良一，「Readability新公式提案に向

けてのプログラム実装」，日本言語テスト学会（JLTA）第18回研究例会，12/20/2003.
40）Nation, I. S. P. *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press, 2001.
41）国際交流基金　日本国際教育協会,『日本語能力試験出題基準（改訂版）』，凡人社，東京，2002.
42）国際交流基金　日本国際教育協会, 前掲書　p.219.
43）川村よし子,「語彙チェッカーを用いた日本語教科書の分析」，*The Second International Conference on Computer Assisted System for Teaching & Learning Japanese CASTEL/J'99 Proceedings*, 1999a, pp.132-137.
44）川村よし子,「語彙チェッカーを用いた読解テキストの分析」,『講座日本語教育』，第34巻，1999 b，pp. 1-22.
45）川村よし子，1999 a，前掲論文.
46）川村よし子，北村達也,「インターネットを活用した読解教材バンクの構築」,『世界の日本語教育（日本語教育事情報告編）』，第6号，国際交流基金日本語国際センター，2001，pp.241-255.
47）国際交流基金 日本国際教育協会，前掲書　p.224.
48）国際交流基金 日本国際教育協会，前掲書　p.224.
49）川村よし子，1999 a，前掲論文.
50）佐野洋，日本語処理プログラム集CLTOOL（第1.2版）http://sano.tufs.ac.jp/cltool，および，佐野洋,『Windows PCによる日本語研究法－Perl, CLTOOLによるテキストデータ処理』，共立出版，東京，2003.
51）宮浦国江,「テクスト・タイプ」,『英文読解のプロセスと指導』（津田塾大学言語文化研究所読解研究グループ編），大修館書店，東京，2002，pp.118-136.
52）宮浦国江，前掲論文.
53）Cote, N., Goldman, S. R., & Saul, E. U., "Students Making Sense of Informational Text: Relations between Processing and Representation", *Discourse Processes*, 25, 1998, pp. 1-53.
54）宮浦国江，前掲論文.
55）堀場裕紀江,「アセスメント」,『英文読解のプロセスと指導』(津田塾大学言語文化研究所読解研究グループ編），大修館書店，東京，2002，pp.243-265.
56）国際交流基金　日本国際教育協会，前掲書，p.52.
57）国際交流基金　日本国際教育協会，前掲書，p.224.
58）Wible, D. et al.，前掲論文.
59）佐野洋,「個人適合の語学教材開発とコーパスの利用について－教育環境の変化に対応する教材開発手法の提案－」，（内山将夫，佐野洋，菅谷史昭，宮田高志，中條清美，西垣知佳子，原田康也.「パネル討論：言語教育・言語学習と知的情報処理研究」，電子情報通信学会　思考と言語研究会）2003年12月.

注

注1）投野（2003）[3]にはgraded readersのコーパス化，海外の小学校レベルの教科書等をコーパスデータとして用いることが提案されている。

注2）見やすいように英語と日本語の出力を段違いにし，downtownに対応する日本語文字部分を太字で示した。

注3）英語テキストの延べ語数はWORDの文字カウントを使用し，日本語の延べ語数はWeb上で公開されている語彙チェッカーを利用した。日本語解析に使用したプログラムの処理能力に限界があるので，日本語テキストの延べ語数は，3.3で作成したサンプルの延べ語数を求め，サンプルの対の数とテキストのすべての対の数から推定した。

注4）散文データについては，Webページの日英対訳文対応付けデータリストより，英語タイトル順のNo. 1からNo. 40までの散文データをCSVファイルとして保存した。そのうち，1編のページ数が15ページ以上のもの（オリジナルタイトル番号：No. 4，11，15，26，32），「詩」など他のテキストと比べて文の長さが短いもの（No. 24），ブラウザ上で対訳が部分的に非表示のもの（No. 20，21），文法的に不正確さがあることを示唆する訳注があるもの（No. 27），以上の9編は調査に含めなかった。

注5）日本語テキストファイルのうち19 KB以内のものは1編のテキストデータ全体を1サンプルとし，1編が19 KBを越えるもの（オリジナルタイトル番号：No. 1，5，7，8，12，25，28，33，34，35，36，37）は各テキストの2箇所から日英200文対をランダムにサンプリングし，日英各2ファイルを作成した。タイトル番号No. 38，40は日本語解析プログラムの処理能力に合わせて日英対訳対の数を150対とした。

注6）"Terms like *USA*, abbreviations like *lbs*, numerics like *123*, symbols like ＋, and monetary amounts like *$3.87* – all are treated as words. With this in mind, for readability evaluation purposes, it is normally a good idea to either select sample text that is absent such entries, or text that has had such edited out before using it in Readability". (*Readability Calculations*, p.6)

注7）"a score of 8.0 means that an eighth grader would understand the document" from "Readability and its Implications for Web Content Accessi-

bility,” http://wats.ca/resources/determiningreadability/1

注 8 ）“By using and comparing their results a number of times – with one another, with comprehension test scores, and with your own judgement – you may soon determine which of the formulas is most reliable with your materials”. (*Readability Calculations*, p.2) 中條他（2004 予定）[27]で，英語教科書，大学入試問題の出題英文など 121 サンプルのリーダビリティを調査した結果，Flesch-Kincaid Formula と Fry Graph が安定して信頼できるスコアを表示することが判明している。また，国内外でリーダビリティを調査した先行研究においても，Flesch-Kincaid Formula が多く使用されており，Fry Graph は米国の教育関係者に多く使われているようである。

注 9 ）日本人英語学習者向けのリーダビリティ公式はまだ実用に至っていないため，米国の英語母語話者向けのリーダビリティを使用した。最近，日本人英語学習者のためのリーダビリティ公式が新井 (2003)[38]，宮崎他 (2003)[39]によって提案され，実用化が期待される。

注 10）WORD の文字カウントでは英文字や数字は数えないが，句読点は 1 文字と数えているので，実際の漢字含有率はもう少し高めになる。最近，佐野 (2003)[50]により詳細な漢字分析が可能な日本語処理ソフトウェアが公開されたので，今後は漢字含有率の測定精度を向上させることが可能となった。

注 11）本稿では 8.9 年生を日本人英語学習者の読書レベルの目安と考えた。教育現場での実情を鑑みると，基準を**図 2** における高 3 の学校英語教科書でなく，高 1 と高 2 の英語教科書におき，その平均である 6.2 年生とした場合の方が現状に近いかもしれない。その場合，6.2 年生以下は 14 編 (42%) となる。

（H 16.1.10 受理）