資　料

日本教育工学会論文誌 33(4), 449-458, 2010

国語科教科書を基にした小・中学校の文章難易学年判定式の構築†

柴崎秀子*1・玉岡賀津雄*2

長岡技術科学大学教育開発系*1・名古屋大学大学院国際言語文化研究科*2

本研究では，小学１年から中学３年までの国語科教科書に収められたテキストの構成要素を分析し，学年による文章の難易尺度を検討した．学年を判定する式を構築するために，主要な国語科教科書の243テキストから，標準的な205のテキストを選んだ．そして，①１文の平均文字数，②１文の平均文節数，③１文の平均述語数，④テキスト全体の漢語の割合，⑤テキスト全体の平仮名の割合の５つを独立変数として，学年を従属変数とする重回帰分析（ステップワイズ）を行った．その結果，平仮名の割合と文の平均述語数の２つが有意な独立変数となり学年についての高い予測力を示した(R^2=0.791)ので，この２変数で学年判定式を構築した．

キーワード：国語科教科書，リーダビリティー，読書教育，形態素解析

1. 研究の背景と目的

2000年以降３年ごとに調査が行われている「OECD生徒の学習到達度調査」(Programme for International Student Assessment，略称 PISA) の読解問題における日本の順位は，2000年に32か国中８位，2003年に41か国中14位，2006年に56か国中15位と下降傾向にある．得点でみても，2003年と2006年はともに498点という成績で，平均点の500点よりも低い結果であった．読解力をつけるには，まず「読むこと」自体が効果的な方法だと言われている（山崎 2008; ATWELL 1998）が，全国学校図書協議会の調査[1]によると，1978年から2008年までの小中高生の１ヶ月の平均読書量は増減を繰り返しながらも徐々に増加しており，2008年の小学生の読書量は1978年のそれの約３倍にもなっている．読書量に深刻な問題がないとすれば，何を基準にどのような本を読むかということが検討課題となろう．

海外には書籍を選択する指標として，米国のレクサイル[2]とリード指数[3]というシステムがある．米国の小中学生は読解力診断テストの結果を元に，レクサイルホームページにある５万冊のデータベースから各自の読解力に適した本を検索できるようになっている．指数５が最も易しく，2,000が最も難しいという設定で，例えば，J.K.ローリング著『ハリーポッター』のシリーズは，指数880から1,030の間にあると計算されている．レクサイルは米国の小中学校や図書館で利用されているだけでなく，大手書店には販売書籍に題名，著者名，価格とともにレクサイル指数を表示しているところもある．また韓国でも，小中学生が読解力診断テストの結果を基に，リード指数を使って読解力に即した本を探すことができるようになっている．リード指数は100が最も易しく，1,850が最も難しいという設定で，例えば，ダン・ブラウン著『ダ・ヴィンチ・コード』の韓国語翻訳版は970である．「OECD生徒の学習到達度調査」の読解力テストで，韓国は2000年に６位，2003年に２位，2006年では１位と徐々に順位を上げてきたが，このリード指数の活用も読解力向上の一因と見ることができるのではないだろうか．

我が国でも2008年６月６日の参議院本会議において2010年を「国民読書年」とする決議が採択され，読書運動への関心が高まっていることが窺える．『何をどう読ませるか』（全国学校図書協議会編）には，「子どもの読解能力等にもっとも適した読書材が選ばれなければならない」とあるが，これは読書教育を考える人々

2009年4月22日受理

† Hideko SHIBASAKI*1 and Katsuo TAMAOKA*2 : Constructing a Formula to Predict School Grades 1-9 based on Japanese Language School Textbooks

*1 Nagaoka University of Technology, 1603-1, Kamitomioka, Nagaoka, Niigata, 940-2188 Japan

*2 Graduate School of Languages and Cultures, Naogya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi-ken, 464-8601 Japan

の共通の意識であろう．小中学生向けの本をレベル別に紹介したものとしては，前述の『何をどう読ませるか』以外に『どの本読もうかな？1900冊』(日本子どもの本研究会編)，『子どもにすすめたいノンフィクション』(日本子どもの本研究会ノンフィクション部会編)，『親も子も読む○年生の読み物全６巻』(学校図書編)，『国語力読解力がつく教科書文庫１年から６年』(川北亮司編集) など多数あるが，いずれも教育者や学識経験者による教育的配慮と経験測に基づくものであり，現在のところ，客観的な数値で示すような指標はない．

そこで，本研究では小中学生の図書選択基準の土台となるような日本語テキストの難易を数値で示す尺度を構築したいと考えた．まず，国語の教科書でデータベースを作り，テキストの構成要素（文字，単語，文の長さ，文法など）を取り出し，学年を差別化できる変数を見つける．次に，これらの変数を独立変数とし，小中学校の各学年を判定する式を重回帰式によって作る．最後に，本研究で構築された式でいくつかのテキストを測定し，その結果を他の文章難易測定ツールと比較し検討する．以上が本稿の流れである．

2. 学年を区別するテキスト構成要素

日本語のテキストの構成要素としてはどのようなものが考えられるだろうか．

第一に，認知心理学の分野では，文は長いほど認知処理に負荷がかかることがわかっているので，文の長さは妥当な変数であろう．英語のリーダビリティー公式の多くが，変数を１文の平均単語数と１単語の平均音節数を変数としている[4]が，これは「長さ」を変数としているからである．英語は単語と単語の間にスペースを置く決まりになっているため，語数および文字数を数えることが容易である．しかし，日本語の場合，次章で述べるように「長さ」を決定するだけでも様々な困難があり，英語のように簡単ではない．本研究では日本語の言語的特徴を視野に入れ「長さ」を考える．

第二に考えられる要素は，日本語には漢字，平仮名，カタカナ，ローマ字の４種類の文字があるという点である．１つの言語に４種類もの文字があることは，他の言語に見られない特徴であり，この点は無視できないだろう．建石ほか (1988) では連続する同一文字種の相対頻度を変数として公式を作っており．また，SATO *et al.* (2008) でも文字の出現確率からモデルを作っている．いずれも日本語に複数の文字種があることを生かした方法である．本研究でも４種類の文字種の割合が難易度を決める変数の１つであると想定した．

第三に，日本語の語彙には和語，漢語，外来語，混種語など複数の語種があるが，近い意味を持つ単語でも，和語の「危ない」は比較的低い学年のテキストに出現し，漢語の「危険」は比較的高い学年のテキストに見られる．このことから，高い学年には漢語が多いことが予測されるので，語種の割合も学年を区別する可能性があると考えられる．

第四に，文法構造の複雑さもテキストの重要な構成要素の１つだと考えられるが，現在までに200以上もあるリーダビリティー公式に文法構造を変数としたものはない．このことは，文法構造を変数とすることの難しさを示唆するものであるが，述語を１つしか持たない単文よりも，複数の述語を持つ複文のほうが難しいと予測される．そこで，本研究では１文あたりの述語の数を文法的な複雑さを示す変数とした．述語の数は，同時に命題の数とも言える．研究者によって定義が若干異なるが，命題は基本的に１つの項と１つの述部から成る単位で，認知心理学の分野では，「文字数が同じ文ならば，命題が多い文のほうが読む速度は遅い」(GOETZ *et al.* 1981) という実験結果がある．この意味からも述語の数は有効な変数となろう．

以上の４つのテキスト構成要素が本研究の分析対象であるが，これらの他にも，テキストの難易を決定するものはあるだろう．例えば，語彙の使用頻度や親密度も重要な変数であると予測される．語彙に関しては，その特性を調査した『NTT データベースシリーズ・日本語の語彙特性』(天野・近藤 2003) がある．しかし，これは語彙の使用頻度を朝日新聞の記事から算出し，親密度は成人を対象とした調査に基づいて作られているため，発達段階にある児童・生徒における語彙の難易度判定には必ずしも適切ではないと思われる．さらに，文章は単なる文字や文の集合体ではなく，接続，結束性，照応などの変数も考えられる．しかし，言語処理技術には限界があり，すべての言語要素を分析できるわけではない．例えば，日本語の文法には係り受け構造があり，並列関係と連体修飾関係を比べた場合，直感的には後者のほうが難しいと思われるが，係り受け解析ツール CaboCha (KUDOH and MATSUMOTO 2000) は係り受けの量は数えられても，両者を区別して数えることはできない．これが区別できれば，それぞれの係り受け関係の数は強力な変数となることが予想されるが，現段階の技術ではそこまでは望めない．接続，結束性，照応についても同様である．本研究ではテキ

低学年よりも全文字数における漢字の割合が大きいことが予測される．そこで，テキストごとに，漢字，平仮名，カタカナ，ローマ字の数を数え，それぞれの文字の総数に対する割合を求めた．その結果，**表２**に示したように，学年が上がるにつれて平仮名が減り，反対に漢字が増えていくことがわかった．また，片仮名とローマ字の割合はどの学年においても極めて小さく，学年による差もほとんどない．この結果から，平仮名または漢字の割合のいずれかを変数にできるが，低学年のテキストには漢字が全く使われていないテキストもあるので，平仮名の割合を変数として使った．

3.2.3. 語種の割合

テキストに出現する全単語は，語種辞書の「かたりぐさ」で判定した．和漢語，和外語，漢外語などもあるが，語数が僅少であったので，和語，漢語，外来語，混種語のみを分析対象とした．学年が上がるにつれて和語は減少し，対照的に漢語は増加する．また，外来語と混種語は学年による変化がほとんどない．語種は，文字種の割合を反映しているとも言えるが，漢字表記の和語も多いので，文字種とは区別して語種の割合も変数として加えるのが適当であると考えた．語種の割合を示す変数としては，和語と漢語のどちらも使えるが，本研究は漢語の割合を使うことにした．

3.2.4. 述語の数

本研究では述語を以下のように定義して算出した．

① 出現した全部の動詞

②「形容詞＋名詞」(例：赤い花) の形で出現しない形容詞 (例：空は青く，山は緑だ．父の手は大きい．)

③「形容動詞＋名詞」(例：偉大な仕事)の形で出現しない形容動詞 (その男は正直で，誠実だった．)

④名詞＋判定詞 (例：明日はよい天気でしょう．これは母の鏡だ．次は渋谷ですか．)

⑤名詞＋句点，すなわち体言止め (例：空からふる白いものは雪．)

⑥非自立名詞＋助動詞 (例：のだ，のです)

①については，２つ以上の動詞から成る複合語は１語と数えた (例：入り込む，連れ出す，呼びつける，走り回る，教えてもらう，来てくれる，歩いていく等)．連体修飾を含まない二重主語文 (例：この町は緑が多い．) は形容詞「多い」が述語としてカウントされるが，連体修飾で終わる二重主語文 (例：ここは緑が多い町だ．) は，形容詞「多い」の直後に名詞「町」があるため，②の定義では除外されるという問題がある．しかし，コンピュータをカウントツールとして使うには何らかの定義を与えなければならないので，この場合は後者を捨てることにした．もし，後者を入れると

表２　243サンプルの学年ごとのテキスト構成要素の平均値

学年	1文の長さ		文字種の割合(単位：%)				語種の割合(単位：%)				文法の複雑さ
	文字数	文節数	平仮名	漢字	片仮名	ローマ字	和語	漢語	外来語	混種語	述語数
1	19.61	4.18	94.15	2.33	2.59	0.00	91.73	6.20	1.44	0.63	1.51
2	20.60	4.72	84.24	9.40	4.88	0.00	86.21	10.63	1.91	1.25	1.56
3	26.49	5.73	79.85	15.04	3.65	0.00	83.20	13.50	1.50	1.80	2.16
4	26.93	5.70	75.39	17.87	5.30	0.00	79.70	16.80	1.90	1.60	3.16
5	29.17	6.18	73.83	19.50	5.08	0.04	76.60	20.00	1.80	1.50	3.37
6	29.08	6.41	69.79	24.32	4.57	0.00	72.70	23.50	2.30	1.50	3.29
7	30.62	6.98	68.00	26.50	4.84	0.03	69.17	26.40	2.28	2.15	3.57
8	30.07	6.79	65.10	27.41	5.69	0.10	66.41	28.85	2.53	2.22	3.35
9	33.16	7.48	65.10	27.86	5.55	0.03	65.44	29.43	3.21	1.92	3.91

表３　205サンプルの各変数における平均と標準偏差

学年	*n*	1文の文字数		1文の文節数		平仮名の割合(%)		漢語の割合(%)		1文の述語数	
		平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差	平均	標準偏差
1	25	20.88	4.29	4.43	1.17	94.80	4.75	5.86	3.62	1.56	0.32
2	25	21.79	6.24	4.92	1.91	84.24	5.32	8.11	5.00	1.61	0.40
3	25	29.43	7.92	6.56	2.32	80.08	5.07	15.11	5.81	2.42	0.66
4	13	26.90	5.32	6.54	4.41	79.54	4.14	14.78	5.86	3.10	0.59
5	20	31.17	7.67	6.77	2.12	75.70	6.33	19.14	5.38	3.62	0.97
6	28	32.01	8.41	7.13	1.94	69.79	6.43	24.84	11.45	3.59	0.85
7	31	35.28	8.17	7.99	1.86	67.71	5.63	26.59	9.64	4.00	0.78
8	25	35.23	10.12	8.08	2.46	63.76	5.09	30.02	9.13	3.82	1.07
9	13	42.12	9.16	9.39	2.22	59.92	3.87	38.02	8.80	4.61	0.85

注：*n*=学年ごとのテキストの数

ストの構成要素として妥当であると考えられ，かつ技術的に分析可能な上記の４つの要素を用いて，学年を予測する式を算出することにした．

3. 方　　法

3.1. テキストの選出

テキストのデータベースとして，光村図書・東京書籍・教育出版の小学１年から６年までの国語の教科書（上巻・下巻）の合計36冊を，また光村図書・東京書籍・三省堂の中学１年から３年までの国語科の教科書の合計９冊を使った．国語科教科書は，漢字，語彙，作文，話し合いなど学習目的によって様々な教材があるので，まず，読解を目的とする散文の教材のみを選び，詩，短歌，俳句などの韻文は排除した．次に，複数の教科書に掲載されているテキスト（例：小学１年「おおきなかぶ」，小学３年「モチモチの木」，小学５年「大造じいさんとがん」，中学１年「走れメロス」，中学３年「故郷」など）は重複を避け，１つだけを使うことにした．対象テキストを OCR で電子化した後，人手で１行ずつ本文と照合しながら誤りを修正した．その結果，テキスト数243，文字数575,215，文の総数19,990のコーパスが完成した．なお，本研究で作成したコーパスは題名，著者名，注，解説文などは除外してある．サンプリングの結果を**表１**に示した．

3.2. テキストの難易度を決める変数

小学１年から中学３年までを１から９の数値で示す従属変数として，学年を予測する重回帰分析を行う．テキストの難易度を決める独立変数は，次の４要素における５変数とした．

3.2.1. 文の長さ

本研究では文の単位を句点で区切れる単位とした．すなわち，句点の数がそのまま文の数を反映し，かぎかっこの中の句点も地の文の句点と同様に数えた．１文の長さは文字数と単語数に反映し，英語のリーダビリティー公式では，１文の平均文字数か平均単語数のいずれか，あるいは両方を変数とするものが多い．しかし，文字種がアルファベット１種類の英語と異なり，日本語には複雑な問題がある．日本語の文字には漢字，平仮名，片仮名，ローマ字の４種類があり，同じ内容の文でも①「教養／課程／が／開設／される」は10文字だが，②「きょうよう／かてい／が／かいせつ／される」は16文字である．柴崎・沢井（2007）では，この問題を解決するために，文の長さを決める変数として，１文の文字数と単語数の両方を使っている．その理由は，①と②の文字数は異なるが，単語数は斜線で区切ったようにどちらも同じで５語となるからである．しかし，柴崎（2008）は形態素解析ツール ChaSen（松本 2000）による複合語の解析が完璧ではないという問題を指摘している．本研究での分析は2007年に行われたため，ChaSen2.3.3が使われたが，例えば，小学３年生の教材「手ぶくろを買いに」のテキストに出てくる「母さんぎつね」は，「母さん(名詞)ぎ(未知語)つね(名詞-固有名詞-人名)」と解析され，テキストに出現する1,174語の内容語の中で602語が未知語として解析された．「母さん」「きつね」「狐」という解析はできるが，「ぎつね」という濁音化した複合語がないため，未知語として認識されてしまうのである．最近では中単位，長単位の解析もできる方法（富士池ほか 2008）も発表され，近い将来，こうした複合語の問題も解決されるであろう．しかし，現段階ではそれは望めないので，以下のように考えた．文字数はテキスト全体の量にも関係するので，これは変数として考える．そして，単語数は変数とせず，文節数を使う．なぜなら，文節は係り受け解析ツール CaboCha で数えることが可能で，ほとんどの文節が内容語と助詞の組み合わせなので，認識しにくい複合語が出現しても，単語の直後に助詞がなければ文節としてはカウントされないからである．以上の考えに基づき，文字の定義を平仮名，漢字，片仮名，ローマ字として，テキストごとに文字数を算出し，文の数で割った．同じく，CaboCha 0.60pre4で文節の数を算出し，文の数で割った．**表２**は各学年の平均値であるが，学年が上がるに従って１文の平均文字数も平均文節数も増えていき，文が長くなっていくことが示された．

表１　国語教科書サンプリング

学年	テキストの数	文字数	文の数
1	27	15,829	807
2	25	30,478	1,454
3	26	41,514	1,567
4	23	41,715	1,549
5	24	65,115	2,232
6	28	84,058	2,891
7	32	96,134	3,140
8	29	99,239	3,300
9	29	101,133	3,050
合計	243	575,215	19,990

3.2.2. 文字種の割合

国語科教科書には学年配当漢字があり，学年が上がるにつれて配当漢字は累積されていくので，高学年は

および他の４つの独立変数と強い負の相関を示し，学年が上がるにつれて平仮名の割合が低くなる傾向が(r=−0.867, p<.001)顕著にみられた．また，漢字の割合が多くなれば，当然のことながら，平仮名の割合は少なくなる(r=−0.870, p<.001)．１文の文字数は，文節数(r=0.881, p<.001)および述語数(r=0.892, p<.001)と強い正の相関を示しており，文字数が多くなれば文節数および述語数も多くなることがわかる．

3.3.3. 重回帰分析による学年判定式の算出

205のテキストの５つの構成要素を独立変数とし，学年を従属変数として予測する重回帰分析をステップワイズ法によって行った．その結果，漢語の割合が除外され，他の４変数が有意な独立変数として学年を予測した．しかし，変数間の相関係数が極めて高いことから，多重共線性を示す可能性が疑われた．そこで，各変数の重相関係数を観察したところ，①１文の平均文字数の許容度は0.204，②１文の平均文節数の許容度は0.376と低い数値が示された．そのため，この２変数を除外し，最終的に文章中の平仮名の割合と１文の平均述語数の２変数を独立変数として，学年を予測する重回帰分析を行った．その結果，平仮名の割合(β=−0.145, p<.001）と１文の平均述語数（β=0.587, p<.001）の標準偏回帰係数はともに高く，有意に学年を予測していることがわかった．重回帰分析の結果を**表５**に示した．重回帰式は以下の通りである．

$Y=-0.145X_1+0.587X_2+14.016$

Y=学年

X_1=文章の総文字数に対する平仮名の割合

X_2=１文の平均述語数

この式の決定係数は，R^2=0.791（調整済み R^2=0.789）と高く，わずかに２つの独立変数であるが，この２変数で学年を良く予測できることがわかる．

4. 考　　察

日本語リーダビリティーを実用化したツールとしては，これまでに以下の２件がある．１つは建石ほか（1988）で発表された公式をテキサス州立大学のAccessibility Institute[5]のホームページで公開したツールである．式の変数は以下に示すように①１文の平均長，②文字種（漢字，平仮名，片仮名，ローマ字）の中の連続する同一文字種の相対頻度，③文字種ごとの連続の平均の長さ，④読点の数の句点の数に対する比である．

−(0.12*LS)−(1.37*LA)+(7.4*LH)−(23.18*LC)−(4.67*CP)+115.79

LS = 文の長さ

LA = 各 run 中の標準的なローマ字と記号の数

LH = 各 run 中の標準的なひらがなの数

LC = 各 run 中の標準的な漢字の数

LK = 各 run 中の標準的な片仮名の数

CP = 読点と句点の割合

Run = 同じ種類の文字が継続的に書かれている事

この式で測定される数値は学年ではなく，ゼロから100までの数値であり，ゼロに近づくほど難しく，100に近づくほど易しいということになっている．建石らは大学生と大学院生を被験者としたクローズ法による実験で式の妥当性を得たとしているが，この式は文字種のどれか１つが極端に多い文章の難易度測定には不向きであり，「自然な表記で書かれたものにしか適用できない」（建石ほか 1998；p.6）と述べている．しかし，ここで言う「自然な表記」とは具体的にどのような表記を意味するのかわからない．試しに，小学１年から高校３年までの国語教科書に収録されているテキストを無作為に選び，建石の公式で測定したところ，「じどう車くらべ」（小学１年）「お手紙」「頭のかきの木」（小学２年）など平仮名の連続する低学年教材の測定値はマイナス１という結果であり測定できなかった．

もう１つはSATO *et al.* (2008) によるもので，「ことば不思議箱」[6]としてインターネットで公開されている．これは小学１年から大学までの13レベルの文字生起確率モデルから成るもので，当該テキストに対して，この13レベルの尤度を難易度順にプロッティングし，最適な推定値を出力する．コーパスは小学校から大学までの教科書127冊から抽出した1,478サンプル（約100万字の教科書コーパス）が用いられ，難易度は１（小学１年）から13（大学）までの値で示される．このツールを使って「じどう車くらべ」（小学１年）「お手紙」「頭のかきの木」（小学２年）を測定したところ，いずれも正確な学年を予測した．さらに，他の小学１年と２年の教材20篇を無作為に選んで測定したところ，どれも当該テキストを収録する学年を示した．小学１年の学年配当漢字は160字，２年は200字と極めて限定された文字しか使用されていないため，低学年用テキストにおける学年判定は，同一文字の連続性を変数とする建石らの式よりも，文字の出現確率を土台とする「ことば不思議箱」のほうが優れていると言えよう．

それでは，本研究で得られた式と「ことば不思議箱」とを比較して検討してみることにする．**図１**の実学年

したら，「大きい町がある.」のような連体修飾をすべて入れることになり，膨大なカウントミスが予想されるからである．③についても同様である．④については形態素解析に判定詞という定義がないので，名詞＋助動詞という定義で数えた．形態素解析は当初ChaSen2.3.3で行っていたが，途中から MeCab0.97とIPA 辞書 2.7.0に変更した．しかし，平仮名の多い小学校低学年のテキスト分析が正確に出来ないため，さらに MeCab0.97と UniDic1.3.9に変更した．しかし，それでもなお，「かさこじぞう」（小学２年）のような方言や口語表現も多いテキストは解析が十分にできないので，小学１年と２年の教材は手作業で述語を数えた．その結果，学年が上がるにつれて緩やかな増加の傾向が見られた（**表３**）．

3.3. 分　析

国語の教科書のテキストを構成する要素を分析した結果，(1) 文の長さは①１文の平均文字数，及び②１文の平均文節数，(2) 文字種の割合は③文章中の平仮名の割合，(3) 語種の割合は④文章中の漢語の割合，(4) 文法構造の複雑さ，及び命題の数は⑤１文の平均述語数の５変数を取り出した．各変数の分布を観察したところ，①１文の平均文字数，②１文の平均文節数，④文章中の漢語の割合，⑤１文の平均述語数の４変数は，学年が上がるにつれて大きくなり，反対に，③文章中の平仮名の割合は学年が上がるにつれて小さくなる傾向にあった．中学１年では，①１文の平均文字数，②１文の平均文節数，⑤１文の平均述語数がやや多くなる傾向にあるが，学年をＸ軸にとった場合，全体的に右肩上がりに変化する傾向を示した．この結果から，これらの５変数を独立変数として，１から９で示した学年を従属変数として予測することが可能であると判断されるので，線形の重回帰分析により学年判定式を算出することにした．

3.3.1. 外れ値の判定と最終的な分析用サンプル

重回帰分析を適用する際，変数が他のテキストの変数と大きく逸脱するものは外れ値として除外する必要がある．例えば，「じゅげむじゅげむ」（小学３年）は１文の文字数が極端に多い．このような特殊なテキストを除外するために全243テキストの①１文の平均文字数，②１文の平均文節数，③文章中の平仮名の割合，④文章中の漢語の割合，⑤１文の平均述語数を独立変数，学年を従属変数として重回帰分析（強制投入法）を行い，予測値を出した．その結果，予測された学年が実際の学年と±2.0以上の誤差のあるデータが38あったので，これを外れ値として除き，205のテキストを最終的な分析のために残すことにした．１から９までの学年ごとの変数の平均は，**表３**に示した通りである．

3.3.2. 学年とテキストの特徴を示す５変数の相関

205のテキストについて学年とテキストの特徴を示す５変数のピアソンの相関係数は**表４**に示した通りである．全体的に相関係数は高く，すべての相関係数が0.001％レベルで有意であった．平仮名の割合が，学年

表４　205サンプルの５変数間における相関

#		1	2	3	4	5	6
1	学年	−					
2	平仮名の割合(%)	-0.867***	−				
3	漢語の割合(%)	0.749***	-0.870***	−			
4	１文の文字数	0.575***	-0.657***	0.686***	−		
5	１文の文節数	0.487***	-0.587***	0.619***	0.881***	−	
6	１文の述語数	0.764***	-0.724***	0.692***	0.892***	0.790***	−
	平均値	4.90	75.38	20.14	30.41	6.86	3.10
	標準偏差	2.54	11.56	12.00	9.68	2.61	1.24

注：n=205, ***p<.001

表５　テキストの学年を予測する重回帰分析の結果

	変数名	β	t値	有意確率
X_1	文章中の平仮名の割合	-0.145	-14.133	p<.001
X_2	１文の平均述語の数	0.587	6.150	p<.001
	重決定係数		R^2=.791	

注: 表の重回帰分析はステップワイズ法による．n=205. βは標準偏回帰係数．

うか．いくつかの一般に知られた作品をこの２つの方法で判定してみた．2008年にベストセラーとなった『ホームレス中学生』（田村裕著）は本研究の式では小学６年，「ことば不思議箱」では中学２年，最近若者の間で人気のある携帯小説の『クリアネス』（十和著）は，本研究の式では中学１年，「ことば不思議箱」は小学６年，同じく携帯小説の『白いジャージ～先生と私～』（reY著）は，本研究の式では中学１年，「ことば不思議箱」では中学２年と，ほぼ同じぐらいのレベルで判定された．ところが，『バカの壁』（養老孟司著）は「ことば不思議箱」がレベル13（大学）と判定したのに対して，本研究の式では小学６年と判定してしまう．新書版が小学生レベルであるとは考えられず，この点において，中学３年を上限と取り，平仮名の割合と述語の数を変数とする本研究の式の弱点があるようだ．

本来，文章難易測定の絶対的な方法というものはあり得ず，１つのツールで万能ということはないだろう．何を変数とするかによってツールの長短があり，成人向け，子供向けなど，テキストの特徴に合わせた使い方があると思われる．芝（1957）も指摘するように，文章の難易尺度とは文章に備わる一定不変の絶対的なものではなく，テキストＡよりテキストＢのほうが易しい，あるいは難しいと判断する相対的な指標にすぎない．しかし，今ある技術を使って，便利でより良いものを作るというところに工学の意義があると言えるのではないか．こうした方法で大量の書籍を数値で分類できれば，読書教育に有益な情報を提供できるものと期待している．さらに，この公式を使ったシステムをコンピュータに実装し，誰でも使えるツールとして公開すれば，例えば，国語のテストを作成するためのテキストを選ぶとき，当該テキストが国語教材の何学年レベルと同等かということが簡単にわかって教育現場には有益であろう．

5. 今後の課題

本研究では国語科教科書をデータベースに，テキストを小学１年から中学３年までの９段階に分類する判定式を構築した．今後は以下の課題に取り組んでいきたいと考えている．

第一に，本研究の学年判定式には形態素解析が必要なので，形態素解析の精度が高くなければならない．形態素解析に当初，ChaSen を使っていたが，最終段階では MeCab と UniDic に切り替え，修正を繰り返している．しかし，現段階でも複合辞（複数の形態素がまとまって機能語となるもの）の解析は困難で，例えば「にかんして」は「に」と「かんして」に分けられてしまう．しかし，佐藤ら（2007）による複合辞用例データベース MUST も発表され，この問題も将来解決できると期待している．

第二に，李・柴崎（2008）にもあるように，語彙の難易は重要な変数となることが予測される．例えば，1953年に翻訳された内藤濯訳の『星の王子さま』は小学２年と判定されるが，作品中の「ウワバミ」「うつし」などの語は，現代の子どもはあまり馴染みがなく，2006年の三田誠広訳や谷川かおる訳のほうが読みやすいことが予想される．『NTT データベースシリーズ・日本語の語彙特性』（天野・近藤 2003）で単語親密度を測定すると，「写し」という表記なら親密度は5.594（7が最高値）という結果であるが，「うつし」「ウワバミ」は該当語がなかった．今後，テキストの学年判定に何らかの方法で語彙の難易を入れていく必要があると考える．

第三に，本研究は中学３年までの学年判定をすることが目的であったが，高校以上をどのように考えるかということも大きな課題の１つである．高校は普通高校，職業高校，高等専門学校など学校によって特色があり，教科書も統一されているわけではない．国語については基本的に，高校１年で国語総合，高校２年で現代文Ⅰ，高校３年で現代文Ⅱを使う高校が多いが，作品によっては，現代文Ⅰと現代文Ⅱの両方に収められているものもあり（例：夏目漱石「こころ」，丸山真男「『である』ことと『する』こと」），学年を差別化するのは難しい．大修館書店と明治書院の国語総合，現代文Ⅰ，現代文Ⅱの合計６冊の教科書から各々24テキストを選んで「ことば不思議箱」で測定したところ，判定された学年は，国語総合が平均値9.54（標準偏差2.52），現代文Ⅰが平均値10.67（標準偏差2.50），現代文Ⅱが平均値は9.88（標準偏差2.71）という結果で３つの学年の差別化はできなかった．高校の国語科教科書には哲学的で抽象的な作品も多く，必ずしも中学レベルの延長上にあるとは言えない．柴崎（2009）は，中学と高校の国語科教科書のテキストを説明的文章と物語文章に分類し，平仮名の割合，漢語の割合，１文の文字数，出現漢字のレベルを分析しているが，高校の国語科教科書は，学年による差よりも文章の種類による差のほうが明確であるかもしれない．

本研究では小中学生を読み手と想定した学年判定式を構築したが，読み手が成人か子供か，日本語母語話

は，各作品が収められた教科書の実際の学年であり，９年（中学３年）が最高学年である．判定学年は，本研究の式と「ことば不思議箱」が判定した最大学年と最小学年の範囲を示したもので，本研究の判定式は９年（中学３年）まで，「ことば不思議箱」は13年（大学）までを判定する．判定範囲が異なるので「ことば不思議箱」の方が基本的に不利になるが，その点も考慮しながら比較してみたい．

まず，判定された全学年の範囲を見ると，「ことば不思議箱」は小学３年以上の全学年において判定範囲を広く取る傾向にある．特に，小学６年を小学３年から高校２年まで，中学２年を小学５年から大学までと判定したのは範囲が広すぎるようだ．本研究の式は中学３年までの９を上限としているため，「ことばの不思議箱」のように高く外れることが無いが，それを考慮しても，小学校，中学校，高校，大学の４つの教育課程のすべてを範囲とするなら，学年を判定したとは言いがたいのではないだろうか．一方，本研究の判定式はどの学年も４学年以内の判定範囲に収まっている．中学１年と２年を差別化することはできなかったが，全体に学年が上がるほど階段を下がるように右に寄っている．

次に，個々の学年でみると，「ことば不思議箱」による判定は小学１年と２年において極めて正確であり，特に小学１年については，すべてのテキストを正確に判定した．小学３年と小学４年については，両者に大きな違いはなさそうであるが，本研究の式は実学年よりもやや高めに，「ことば不思議箱」はやや低めに判定している．小学５年と小学６年になると，「ことば不思議箱」の判定する学年範囲はきわめて広いものとなる．例えば，小学５年の「月夜のみみずく」「だいじょうぶ，だいじょうぶ」を「ことば不思議箱」は小学１年と判定している．本研究による式も，この２作品を実際の学年よりも２学年低く判定した．「月夜のみみずく」は詩的な情感を持つ童話で，「だいじょうぶ」は子どもの心の成長が窺える随想である．どちらも平仮名が多く，文章は易しいが内容が深い教材で，小学校高学年にふさわしいと思われるが，文字のみで判定すると小学１年生レベルということになってしまい，この点が文字の生起確率による判定の弱点であると思われる．小学校の教科書には学年配当漢字が定められていることを考慮すると，小学６年まではもっと正確な判定が期待できるはずであるが，小学５年から判定範囲が広すぎる傾向が見られる．中学の教材についても同様で，中学２年と中学３年の「文化を伝えるチンパンジー」「考えるイルカ」「社会調査のうそ」「メディア・リテラシー」「生物の多様性と環境」「テレビ映像の本質」「テクノロジーとの付き合い方」「テクノロジーと人間らしさ」の８作品を「ことば不思議箱」は大学レベルと判定している．これらの作品には，「埴」「阪」「哉」「沌」「恣」「餌」「藤」「葛」「脆」「惧」など常用漢字外の漢字が使われているために，このような判定になるのではなると思われる．中学以上は漢字の学年配当がないので，文字の生起確率を利用して学年を判定するのは難しい面が出てくるのではないだろうか．以上をまとめると，小学校低学年は文字の生起確率で判断したほうが正確に学年を判定できるが，小学校中学年以上は，本研究の判定式のように述語数を加味したほうがより正確な学年が判定できるようである．

それでは国語科教科書以外のテキストはどうであろ

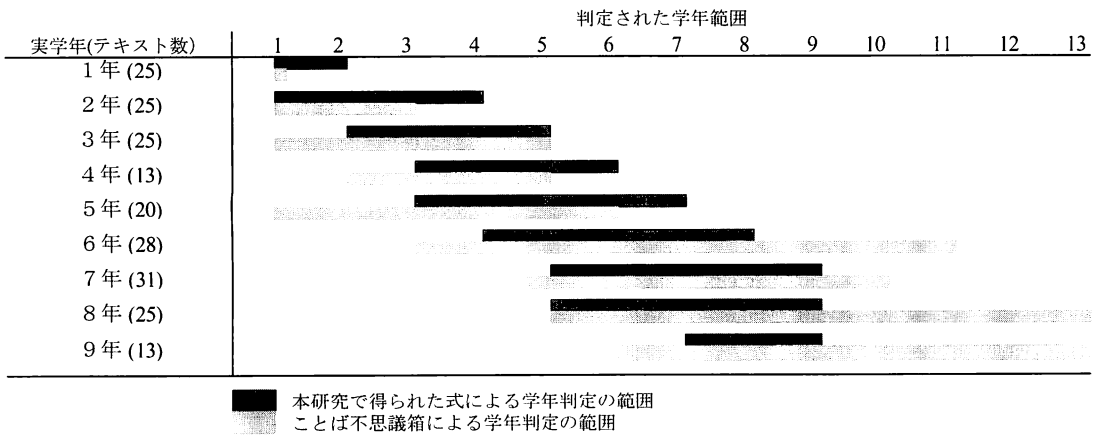

図１　本研究の式と「ことば不思議箱」が判定した学年範囲

Summary

The present study aimed to build a readability formula to measure levels of school grade1–9 based on school texts for the Japanese language. Five predicting variables were selected to determine readability as defined by school grades: (1) average number of characters in a sentence, (2) average number of phrasal units in a sentence, (3) average number of predicates in a sentence, (4) ratio of Chinese-originated words in a text, and (5) ratio of hiragana characters in a text. After excluding deviated texts and variables showing multicollinearity, a linear multiple regression analysis indicated two significant predictors of 'average number of predicates in a sentence,' and 'ratio of hiragana characters in a text', showing a high predicting power (R^2=0.791).

KEYWORDS: JAPANESE LANGUAGE SCHOOL TEXTBOOKS, READABILITY, READING EDUCATION, MORPHOLOGICAL ANALYSIS

(Received April 22, 2009)

者か非母語話者か，非母語話者であれば漢字圏か非漢字圏でそれぞれ異なる方法が必要であろう．今後は本研究の成果を発展させ，様々な読み手を想定した尺度を構築したいと考えている．

謝　　辞

本研究は以下の助成金を受けています．平成19年度～平成20年度科学研究費補助金基盤(B)課題番号1930277 研究代表者・柴崎秀子『日本語リーダビリティー測定尺度の構築とソフトウエアへの実用化』

注

1) 全国学校図書協議会「第54回読書調査の結果」http://www.j-sla.or.jp/material/research/54-1.html 出版物としては，毎日新聞社 (2009)『2009年度版読書世論調査－第62回読書世論調査，第54回学校読書調査』にまとめられている．
2) レクサイル http://www.lexile.com/
3) リード指数 http://www.kyoboread.com/rRead/main.jsp 2006年７月から2007年６月にかけて，1038人の小学生を対象にリード指数の効果と満足度の調査が行われた．その結果，１年間リード指数を使用した小学生の26%が同学年児童の平均読解力よりも優れた成績を修めたと報告されている．
4) Flesch-Kincaid readability test は米国防衛省が文書作成の基準にするほど著名なものであるが，この公式は以下のように，１文の文字数と１語の音節数の２つが変数とされている．206.835-1.015*１文の平均語数-84.6*１語の平均音節数．詳細は，Flesch Reading Ease Test, Rudolf Flesch (1948) A new readability yardstick, *Journal of Applied Psychology*, **32** : 221-233 にある．
5) 建石の式がツールとして公開されていたが，2008年８月以降2009年８月まで，日本語のみ稼動していない． http://webapps.lib.utexas.edu/TxReadability/app
6) ことば不思議箱.http://kotoba.nuee.nagoya-u.ac.jp/

注:1), 2), 3), 5), 6)における URL の参照日 2009.08.09

参 考 文 献

天野成昭，近藤公久 (2003) NTT データベースシリーズ・日本語の語彙特性 : 第２期 CO-ROM 版．三省堂，東京

ATWELL, N. (1998) *In the middle*. Heinemann, Portsmouth, NH

富士池優美，小椋秀樹，小木曽智信，小磯花絵，内元清貴 (2008) 現代日本語書き言葉均衡コーパスにおける長単位の概要．特定領域研究日本語コーパス平成19年度公開ワークショップ研究成果報告会予稿集 : 51-58

GOETZ, E.T., ANDERSON, R.C. and SCHALLERT, D.L. (1981) The representation of sentences in memory. *Journal of Verbal Learning and Verbal Behavior*, **20** : 369-385

KUDOH, T. and MATSUMOTO, Y. (2000) Japanese Dependency Analysis Based on Support Vector Machines. EMNLP/VLC

李在鎬，柴崎秀子 (2008) 日本語リーダビリティー公式構築のための国語科教科書語彙の分析．計量国語学会第52回年次予稿集 : 16-22

SATO, S., MATSUYOSHI, S. and KONDOH, Y. (2008) *Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus*, LREC-08

佐藤理史，宇津呂武仁，土屋雅稔，松吉俊 (2007) 日本語複合辞用例データベース MUST1.0 説明書

芝祐順 (1957) 読み易さの測り方－クローズ法の日本語への適用－．心理学研究**28** : 67-73

柴崎秀子 (2008) 日本語コーパスを応用した文章の難易測定の研究．特定領域研究日本語コーパス平成19年度公開ワークショップ研究成果報告会予稿集 文科省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21世紀の日本語研究の基盤整備」総括班 : 125-130

柴崎秀子 (2009) 日本語リーダビリティー公式の構築とツールの開発．特定領域研究日本語コーパス平成20年度公開ワークショップ予稿集 : 155-160

柴崎秀子，沢井康孝 (2007) 国語科教科書コーパスを応用した日本語リーダビリティー構築のための基礎研究．信学技報, NLC2007-32 (2007-10) : 19-24

建石由佳，小野芳彦，山田尚勇 (1988) 日本文の読みやすさの評価式．文書処理とニューマンインターフェース, **18**(1) : 1-8

松本裕治 (2000) 形態素解析システム「茶筌」．情報処理, **41**(11) : 1208-1214

山崎博敏 (2008) 学力を高める「朝の読書」：一日10分が奇跡を起こす: 検証された学習効果朝の読書．メディアパル，東京